PAR PUBLIC ADMINISTRATION REVIEW

Jesper Asring Hansen
Aarhus University
Lars Tummers
Utrecht University

A Systematic Review of Field Experiments in Public Administration

Research Article

Abstract: *Field experiments have become popular in public administration. By allowing for the identification of causal effects in realistic settings, field experiments may become central in several research agendas of relevance to the field. Conducting field experiments is difficult and problems often occur along the way. However, researchers new to the method have few resources in public administration to consider the problems that arise when conducting field experiments. This systematic review identifies 42 field experiments in public administration and serves as an introduction to field experiments in public administration. The article discusses how field experiments developed over time and highlights trends in field experimentation in public administration. It then discusses issues to consider when designing field experiments. Among these are costs, practicality, ethics, and validity. Finally, the authors suggest a future research agenda for public administration field experiments.*

Evidence for Practice

- Field experiments have high value for public administration scholars and practitioners, as they may allow for causal inference in real-world settings.
- Field experiments are often difficult to conduct.
- Researchers and practitioners conducting field experiments should consider, among other things, the costs, practicality, ethics, and validity of their studies.

Jesper Asring Hansen is assistant professor in the Department of Political Science and Trygfonden's Center for Child Research, Aarhus University. His research focuses on performance management, behavioral public administration, and quantitative text analysis.
E-mail: jesper@ps.au.dk

Lars Tummers is Chair of Public Administration and Organizational Science in the School of Governance at Utrecht University, the Netherlands. His main research interests are public management, stereotypes, leadership, and behavior change. He is developing—with others—an interdisciplinary field combining psychology and public administration, called behavioral public administration.
E-mail: l.g.tummers@uu.nl

Field experiments have become popular in public administration (James, John, and Moseley 2017). Scholars examining topics such as coproduction (Jakobsen 2013), discrimination (Grohs, Adam, and Knill 2016), and leadership (Bellé 2014) use this method. The credibility revolution in economics (Angrist and Pischke 2010) has diffused into public administration, where a focus on the identification of causal effects, combined with preferences for realistic treatments, explains the proliferation of field experiments. Experimental realism is a benefit distinguishing field experiments from other methods with internal validity, such as survey experiments and lab experiments (Baekgaard et al. 2015).

Field experiments might become central to new research agendas in public administration. An emerging literature examines research questions in public administration by combining insights from psychology with public administration (a research paradigm labeled behavioral public administration; see Battaglio et al. 2019; Grimmelikhuijsen et al. 2017). Since many of the central questions of this research agenda (and of public administration in general) relate to the behavior of citizens, civil servants, and politicians, field experiments may shed light on critical questions that remain neglected in public administration. Yet, to date, there are not many field experiments in public administration, though the method is gaining popularity (James, John, and Moseley 2017). We suspect that this may be because field experiments are difficult to conduct, and there is very little inquiry in public administration to guide researchers in handling the difficulties that arise when conducting field experiments. This systematic review serves as an introduction to field experiments in public administration by highlighting topics discussed and questions to consider when conducting field experiments. We provide examples from public administration and related literatures to create a detailed resource on how to conduct field experiments.

We conducted a systematic review of field experiments in public administration to examine field experimental research. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, which provide a 27-item checklist for reporting systematic reviews. Following PRISMA leads to more transparent and replicable studies (Moher et al. 2009). The first part

Public Administration Review, Vol. 9999, Iss. 9999, pp. 1–11. © 2020 The Authors. *Public Administration Review* published by Wiley Periodicals, Inc. on behalf of The American Society for Public Administration.
DOI: 10.1111/puar.13181.

of the analysis considers trends and topics in field experiments. It demonstrates that field experiments are becoming more popular in public administration and that performance, leadership, and coproduction are the most studied topics. After the exploratory analysis, we discuss whether the studies contain "fieldness" and how to distinguish a field experiment from a survey experiment. Fieldness is "the extent to which an experiment is conducted in a true field" (Eden 2017, 97). There are fuzzy boundaries between field experiments and survey experiments, but we find that public administration field experiments fare well in terms of fieldness.

The second part of the analysis considers the costs, practicality, ethics, and validity of field experiments, since most field experimenters have to consider these issues. Field experiments typically demand higher efforts of researchers than other types of studies, which is why it is useful to consider how public administration researchers manage these issues. Furthermore, we highlight different literatures that inform about how to administer these issues appropriately, so that field experiments confer less costs on part of researchers and participants and conclusions are as valid as possible.

This article proceeds as follows: We provide a description of what a field experiment is, after which we describe how we conducted the systematic review and searched for studies. After that, we move on to the analysis; the first part examines the trends and topics studied, while the second part discusses costs, practicality, ethics, and validity. We conclude with a discussion on the future of field experimentation in public administration.

Field Experiments and Public Administration

Field experiments assign units randomly to treatment and control groups, like other types of experiments. They differ from other types of experiments in terms of fieldness (Eden 2017). Fieldness is determined by at least four criteria: (1) the intervention is realistic, (2) participants encounter the treatment in the real world, (3) the context is natural, and (4) outcome measures mirror the outcome of interest (Eden 2017). Czibor, Jimenez-Gomez, and List (2019) propose additional criteria: whether the population consists of students or the population of interest, whether the environment is artifactual or natural, and whether the treatment is overt or covert (Czibor, Jimenez-Gomez, and List 2019). These criteria create four types of experiments: lab experiments, artifactual field experiments, framed field experiments, and natural field experiments (Czibor, Jimenez-Gomez, and List 2019).

The fieldness criteria highlight differences between field experiments and survey experiments. Since survey experiments are embedded in surveys, they typically cannot have treatments similar to what is encountered in the real world. Their contexts are surveys, which are not similar to real-life situations, and the outcomes are often questions (though some surveys include semibehavioral indicators). The many differences between surveys and realistic settings make it unclear how their conclusions map onto behavior in the real world (Levitt and List 2007). The same goes for lab experiments. Lab experiments recruit participants who are assigned to different types of treatments in the lab setting. There are often profound differences between the lab setting and the real world, though lab experiments offer benefits in terms of experimental control (Falk and Heckman 2009). Hence, field experiments allow researchers to gain knowledge of how a causal effect of interest manifests in the real world.

A field experiment in public administration occurs in a natural setting (e.g., a municipality), includes relevant subjects (e.g., civil servants), uses authentic treatments (e.g., citizen demographics), and measures real-world outcomes (e.g., quality of responses). An example that meets all criteria is the study by Jilke, van Dooren, and Rys (2018). This field experiment showed how public and private elderly care homes in Belgium responded to emails from natives relative to Maghrebian immigrants. The scholars were thereby able to compare responses and demonstrate whether discrimination based on ethnicity occurs at different rates in public and private organizations. It is hard to imagine that survey or lab experiments could provide the realism needed to make conclusions about discrimination by public and private service providers.

A Systematic Review of Field Experiments

We conducted a systematic review to assess field experimentation in public administration. We report the PRISMA checklist and how we adhere to the items in the Supporting Information online. There, we elaborate on the articles included in the review, the year they were published, and the topics they cover. All data can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LF9H7Q.

Eligibility Criteria

Studies were included in the systematic review if they met all of the following criteria:

- Study design: Studies should identify as field experiments or as audit studies. They should contain the words "field experiment," "randomized controlled trial," or "audit study" in the title or abstract.
- Topic and type of participants: The topic should be of relevance to public administration. Relevant topics include citizen-state interactions, interventions in public organizations, and the like. Participants in the study should be of interest to public administration. Participants are, for instance, politicians, civil servants, public employees, or citizens.
- Language: We only included studies reported in English.
- Publication status: We only included peer-reviewed journal articles published in public administration journals.
- Year of publication: We included all years of publication to find studies, until 2018 (no restriction). We only found studies published from 2010 to 2018 and some studies that were forthcoming in 2018, and thus published in 2019.

Search Strategy

We used three search strategies. First, we searched electronic databases for the terms "field experiment," "audit study," and "public" [search string: "field experiment" OR "audit study" AND "public"]. We used this search string in Scopus (1,040 results). Second, we searched leading and specialized public administration journals using the search term "field experiment." The journals we searched were *Journal of Public Administration Research and Theory* (48 potential studies), *Public Administration Review* (28), *Journal of Policy Analysis and Management* (25), *Review of Public Personnel Administration* (14), *International Public Management Journal* (14),

Public Administration (9), *Public Management Review* (4), *American Review of Public Administration* (4), *Governance* (2), *Journal of Behavioral Public Administration* (2), and *Regulation & Governance* (1). This search resulted in 151 records. Finally, after reading titles and abstracts, we sent out the preliminary list to three experts in field experiments. The experts identified 12 additional records of which we included two in the review.

Textbox 1. Search string used for Scopus.

String used for Scopus: TITLE-ABS-KEY ("field experiment" OR "audit study" AND "public") AND (LIMIT-TO (PUBSTAGE, "final") OR LIMIT-TO (PUBSTAGE, "aip")) AND (LIMIT-TO (ACCESSTYPE[OA]) OR LIMIT-TO (ACCESSTYPE[OTHER])) AND (LIMIT-TO [LANGUAGE, "English"]).

Record Selection

After screening titles and abstracts of all the records identified through our systematic review, we found 42 studies eligible for inclusion. Figure 1 visualizes the selection process.

Coding

We coded studies by their topics covered to examine the questions highlighted by using field experiments. The coding process was partly exploratory, since we included a new category each time an article in the review did not fit any of the themes. We ended up with seven themes (motivation, coproduction, performance, leadership, transparency, discrimination, and other), which are not mutually exclusive. This means that one article can study two topics at the same time. For instance, a study can relate to the literatures on leadership and motivation at the same time (e.g., Bellé 2014).

Results

Characteristics of Field Experiments in Public Administration

The first part of this analysis highlights the characteristics of the identified field experiments. The first result is the increasing trend displayed in figure 2. Before 2010, there were no field experiments in public administration journals. From 2010 on, field experiments have become more popular. The most field experiments have been published in 2018. Public administration journals published 10 field experiments in 2018 (e.g., Andersen and Moynihan 2018; Jilke, van Dooren, and Rys 2018; Porter and Rogowski 2018). The trend highlights the relevance of considering issues that arise when conducting field experiments.

Table 1 shows the topics studied. The most frequently studied topic is performance (e.g., Andersen and Larsen 2016; Andersen and Moynihan 2018), with 19 percent of the published field experiments, followed by leadership, which experienced a surge due to the LEAP project, in which researchers from Aarhus University provided leadership training to public managers to examine the effects of transformational and transactional leadership

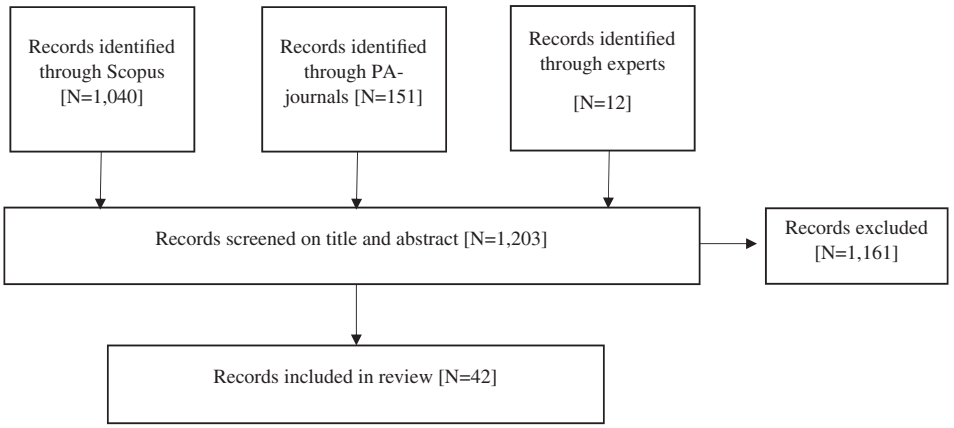

Figure 1 Selection Process

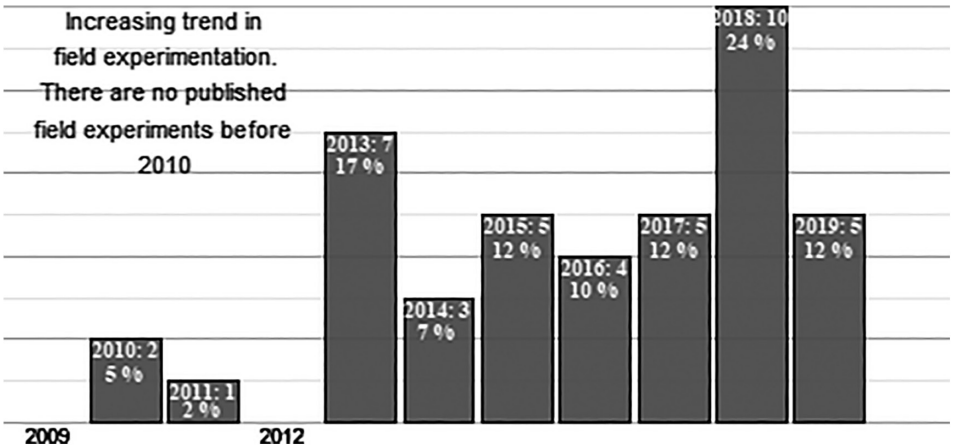

Figure 2 Field Experimentation over Time

Table 1 Topics in Field Experiments

Public Administration Topic	Number/ Percentage	Example Study
Performance	8/19%	Andersen and Moynihan (2016)
Leadership	7/17%	Bellé (2014)
Coproduction	6/14%	Thomsen and Jakobsen (2015)
Transparency	6/14%	Grimmelikhuijsen and Klijn (2015)
Motivation	4/10%	Linos (2018)
Discrimination	4/10%	Jilke, van Dooren, and Rys (2018)
Other	9/21%	Hess, Hanmer, and Nickerson (2016)
Total	44/105%	

Note: Totals may be greater 100%, as some studies combine topics, such as Bellé (2014), who studies motivation and leadership.

on organizational outcomes (e.g., An et al. 2019; Jensen 2018). Another 17 percent of the field experiments identified in this review study leadership, and 14 percent of the field experiments are focused on coproduction (e.g., Jakobsen 2013; Jakobsen and Andersen 2013a). Topics such as discrimination (e.g., Jilke, van Dooren, and Rys 2018), transparency (e.g., Grimmelikhuijsen and Klijn 2015), and motivation (e.g., Bellé 2014) are also popular, but somewhat less so. The category *Other* (covering 21 percent of the public administration field experiments), includes studies on topics such as nudging (John and Blume 2018), government funding and charitable giving (Jilke et al. 2019), and voter registration (Hess, Hanmer, and Nickerson 2016).

Fieldness

The previous section demonstrated an increase in the number of field experiments in public administration research. We can now examine the extent to which the identified studies score on their degree of fieldness. This is important because it shows which fieldness aspects public administration studies already score high on and where potential progress lies. The most clear-cut case of a field experiment is when scholars conduct a study in natural settings with relevant units of analysis subjected to a realistic intervention with the measurement of relevant variables as the outcome. It is not given that a study identified as a field experiment adheres to all criteria. However, most experiments can be placed on a continuum, and the boundary between a field experiment and a survey experiment is not always clear.

Public administration studies identifying as field experiments score high on fieldness when applying the four criteria reviewed in the section defining field experiments (Eden 2017). There are some field experiments in public administration journals that clearly satisfy all criteria (Bellé 2014; Jakobsen 2013; Jilke, Van Dooren, and Rys 2018). These studies assign relevant subjects to experimental groups with realistic treatments, occur in real settings, and have measures of behavior as their outcomes. Other studies satisfy all criteria but have survey questions as outcomes (e.g., Andersen and Larsen 2016; Thomsen and Jakobsen 2015). This increases the similarity to survey experiments, though they still have several features on which they clearly differ from survey experiments. Among these are the fact that they occur in real contexts and have realistic treatments. In other studies that measure outcomes with survey questionnaires, treatments are more realistic than in survey experiments, though the treatments bear resemblance to the way treatments are carried out in survey experiments (James 2011; James and Moseley 2014). However, they still differ from survey experiments by occurring in natural settings, which increases their fieldness relative to normal survey experiments. Finally, some studies identify as "artefactual field experiments," which is why it should be expected that not all criteria for fieldness are met and that the study combines elements from survey experiments and field experiments (Bellé, Cantarelli, and Belardinelli 2017).

Hence, it is difficult to clearly delineate what a field experiment is and how it is distinguished from a survey experiment, but field experiments in public administration generally differ from survey experiments, at least in some respects. Researchers conducting field experiments should consider how they increase the fieldness of their studies to make the studies as realistic as possible, thereby benefiting from the primary strength of field experiments. The area most in need of improvement in the public administration literature in this regard is fieldness of the outcome measure. Behavioral indicators are more indicative of real-world behavior than survey questions, which is the reason why public administration field experiments should consider using such indicators more frequently in the future. Examples of outcome measures of behavior in the public administration literature are task performance (Bellé 2014), quality of responses to emails (Jilke, van Dooren, and Rys 2018), and job applications (Linos 2018).

The conclusions we derive about fieldness may depend to some extent on the taxonomy used for evaluating fieldness. The taxonomy developed by Czibor, Jimenez-Gomez, and List (2019) may lead to slightly different conclusions about fieldness than the ones obtained with the criteria used in this article. The fieldness of public administration field experiments is an issue, which may be subject to further discussion in the future.

Considerations in Field Experiments

We now turn to considerations that researchers are confronted with when conducting field experiments. We do so to help researchers navigate how to set up a field experiment, so that we may see more of them in the future. We discuss costs, practicality, ethics, and validity. The first two issues are highlighted to guide researchers who are new to field experiments, because such studies are costly and difficult to implement. By illustrating how to keep costs down (Broockman, Kalla, and Sekhon 2017) and how to consider practicality (List 2011), we hope that scholars may have higher chances of succeeding in conducting public administration field experiments in the future. The issues of ethics and validity are discussed with the purpose of maintaining the relevance of public administration field experiments. Field experiments in public administration need to consider ethics and validity in the future to preserve their credibility, which is one of the primary benefits they offer to the public administration literature (Angrist and Pischke 2010).

Costs

An important difference between field experiments and other types of studies is that field experiments can entail substantially greater costs for the researchers. This is especially true when comparing field experiments with survey experiments, which typically are easy to conduct and entail low costs (Mutz 2011). The costs of field experiments imply that researchers should use cost-benefit analyses to assess how much power they can achieve with a given level of resources. Several practices aid in keeping costs down (e.g.,

Broockman, Kalla, and Sekhon 2017); we discuss some of them in this section.

One method for reducing costs is collaborating with organizations. Organizations may help deliver treatments and thus bear some of the costs associated with the field experiment. However, research-practice collaborations are no panacea, because they may cause difficulties depending on the infrastructure and interests of the organization (Glennerster 2017; Levitt and List 2009). We discuss these difficulties in the section on practicality.

There is variation in the costs of field experiments, even though they generally are expensive. This is illustrated by considering some of the public administration field experiments. At one end of the continuum, we have large-scale organizational interventions, examining how outcomes change when profound changes in organizational practices or procedures occur. The LEAP project (e.g., Jensen 2018; An et al. 2019) is placed at this end of the continuum, since researchers had to ensure collaboration with the Aarhus municipality, agree on the content of leadership training, deliver the treatment to public managers (which consisted of four leadership training sessions), and, finally, keep track of the organizations to examine whether the training changed outcomes. To our knowledge, this study kept costs down by gathering a baseline wave before delivering the treatment and by combining multiple measures into an index (Jensen 2018). Another example of a costly intervention is the study by Andersen and Moynihan (2018). They examine whether socially distinctive newcomers affect team reflection and what teams think of socially distinctive newcomers. The researchers assigned new hires to some of the schools, who participated in the experiment, which again shows the differences between costs of treatments in field experiments and survey experiments. The outcome measures are also combined into indexes in this case, thereby driving costs down.

Not all field experiments are as expensive as the aforementioned studies. An example of a type of field experiment that is not more costly than most other types of studies is the audit study. In audit studies, researchers send out fake requests to examine whether varying background characteristics affect response rates. Differences in response rates contingent on background characteristics indicate discrimination. This means that the studies have to send out emails to experimental subjects, varying background characteristics. Jilke, van Dooren, and Rys (2018) do so in their study of discrimination based on ethnicity among Flemish elderly care providers. They find that response rates are not different for public and private providers but that public providers tend to give responses of higher quality. Thereby, the researchers provide knowledge of public-private differences in a realistic setting.

Audit studies may be costly for research societies and organizations, even though they are inexpensive for the researchers conducting them. Indeed, respondents are unwittingly recruited to participate in audit studies (Pager 2007, 126). This may decrease trust in the scholarly community. Furthermore, audit studies may be costly to the organizations unwittingly participating because of the time it takes to respond to emails. In highly powered studies, this may imply a significant amount of lost work hours. We discuss the ethics of audit studies further in the ethics section.

Practicality

The practical difficulties of conducting field experiments may be an important reason why there have been few field experiments in public administration to this date. In general, field experiments are hard to implement (see, e.g., List 2011). We are not able to comprehend these practical difficulties by reading published studies because they seldom describe the issues they encountered. Furthermore, the published studies are the field experiments that did overcome the barriers to field experimentation. This implies that we cannot turn to the public administration literature on field experiments to gauge the extent of practical difficulties involved in conducting field experiments.

Some field experimenters have provided recommendations on how to overcome practical barriers to field experimentation. Among these are Glennerster, who recommends pilot-testing and collaborating with implementers (Glennerster 2017; Glennerster and Takavarasha 2013). Pilot-testing aids in finding difficulties in implementing the treatment as well as identifying reasons for noncompliance and attrition (Glennerster 2017).

Collaborations make large-scale studies of interventions more feasible and less costly to the researcher. This is relevant for public administration researchers, who have the possibility of collaborating with public and nonprofit organizations when carrying out field experiments. Glennerster suggests that researchers can use the following criteria to determine whether a partnership is feasible: the organization should have sufficient scale, flexibility, technical expertise, local expertise and reputation, low staff turnover, and a desire to know the truth (for a thorough discussion of the criteria, see Glennerster 2017, 178–180). Finding good partnerships is important, since some researchers have encountered problems when cooperating with organizations. Organizations may have interests diverging from the researcher's interests (Levitt and List 2009). Among these issues are the importance of randomization (Eden 2017), studying questions with higher social returns rather than private returns to the organization (Levitt and List 2009), and ensuring that the organization supports strategies for avoiding noncompliance of experimental subjects (Glennerster 2017). Eden (2017) grants several recommendations on how to manage disagreements between researchers and organizations.

Collaborations may help researchers in running large-scale evaluations of programs but comes at a cost of flexibility. Yet Glennerster states that the benefits of self-implementation seldom outweigh the costs (Glennerster 2017, 191).

Another issue related to practicality is control. Field experiments confer less control, especially when compared with lab experiments, in which variables such as repeated interactions can be manipulated and there can be controlled theory testing (Falk and Heckman 2009; Lonati et al. 2018). The field setting provides more realism, but with this realism comes more complexity, since relations between employers and employees may moderate treatment effects (Falk and Heckman 2009). Falk and Heckman (2009) do not view this as an issue for field experiments but rather as a feature highlighting the importance of considering the causal effect of interest before deciding on the method of inquiry.

Ethics

Ethical considerations are important in field experiments. Field experiments are true interventions, which means that they may have an impact on participants' lives. This demonstrates that ethical considerations are important in public administration field experiments.

These considerations first concern the notion that treatments should not make any people worse off. The public administration literature on field experiments adheres to this criterion. None of the field experiments manipulate variables that are expected to make anyone worse off. Rather, treatments encourage potential positive outcomes such as transparency (Grimmelikhuijsen and Klijn 2015), participation (Hock, Anderson, and Potoski 2013), and equipment stimulating coproduction (Thomsen and Jakobsen 2015).

The second ethical consideration is to waste as little time as possible on behalf of participants. Jakobsen (2013) discusses ethics in his study on citizen coproduction, in which he states that it is not ethical to devise a placebo control treatment, even though it would allow for ruling out the hypothesis that the treatment worked because of a placebo effect (Jakobsen 2013). These considerations apply to audit studies as well, especially because participants cannot give their informed consent as they are unwittingly recruited to participate (Pager 2007, 126). Indeed, some audit studies achieve very high statistical power, which may mean that they take up a lot of time on behalf of participants. For instance, Pfaff et al. (2019) examine whether religious bias shapes access to public services. To do so, they sent out fake requests to 45,000 principals in public schools. Assuming that the treatment took two minutes to process on average, this implies that the researchers used 1,500 work hours of principals in public schools.

Researchers need to thoroughly consider the importance of the research question and statistical power when designing audit studies because of the lack of informed consent. An example of these considerations is the study by Grimmelikhuijsen et al. (2019), who made their requests as transparent as possible while mitigating the issue of noncompliance in their study of the effects of freedom of information laws on transparency. We advise researchers to achieve approval of ethical review boards, as Grimmelikhuijsen et al. (2019) did, if there is potential for ethical violations in the study.

Internal Validity

Like other types of studies, field experiments have to consider whether their conclusions are internally and externally valid. Considerations of internal validity generally focus on whether influences other than the treatment alter the outcomes of participants in the study, while external validity relates to whether the findings of a study generalize to different contexts.

The internal validity of field experiments is related to the assumptions of excludability and noninterference. Furthermore, attrition is a challenge to internal validity. The internal validity of a field experiment is compromised if these assumptions are violated. However, since they are defined through alterations of participants' outcomes, it is not possible to directly observe whether they were violated in the experiment. We now turn to discussing how to accommodate assumptions related to internal validity.

Threat 1: Excludability. The first threat to internal validity is violation of the assumption of excludability. This assumption states that the randomization does not affect outcomes through other variables than the reception of the treatment (Angrist and Pischke 2009, 116). If this assumption is not met, the causal effect identified in a study is a combination of the treatment and other variables. The excludability assumption cannot be demonstrated in general, and researchers have to evaluate the assumption "on a case-by-case basis" (Athey and Imbens 2017, 118). Furthermore, the excludability assumption is dependent on how researchers define the treatment. The treatment can be defined either as being assigned to treatment or as receiving the treatment, and the excludability assumption may be violated with the second definition of the treatment but not the first one. This implies that researchers should preregister whether they are interested in treatment assignment or receiving the treatment to avoid researcher flexibility leading to false positives (Simmons, Nelson, and Simonsohn 2011).

Excludability highlights the importance of observing data-generating processes to demonstrate that violations are unlikely. Some scholars note that causal process observations aid in making claims about causality (Brady and Collier 2004). A classic example is John Snow, who studied the outbreak of cholera in London. By combining detailed knowledge about water supplies in individual households with data from a natural experiment, he was able to show that cholera was a waterborne disease, refuting the prevalent theory that it was spread through bad air (miasma) (Dunning 2012, 12). Similarly, detailed knowledge about field experimental conditions and participants' reactions to them may contribute to making claims about the validity of the study.

A second way to increase the plausibility of excludability is to handle treatment and control groups as similarly as possible (Gerber and Green 2012, 41). Scholars may succeed in doing so by double-blinding treatments, which enhances the likelihood of an equal handling of treatment and control groups. We were not able to find any public administration field experiments that explicitly stated that they were double blinded. However, we did find that scholars in the LEAP project sought to standardize their treatments to make the excludability assumption more likely (e.g., An et al. 2019; Jensen 2018). The experimenters agreed on the content and presentation of leadership training before they treated public managers with a training in transformational or transactional leadership (Jensen 2018). Hence, even without double-blinding, it is possible to increase trust that excludability is a viable assumption in a study.

Threat 2: Interference. Interference occurs when experimental units alter each other's outcomes, thereby contaminating the effects of the treatment (Bowers, Frederickson, and Panagopoulos 2013). This creates a bias that may be both upward and downward, meaning that scholars cannot estimate causal effects when interference is present.

Causal process observations may aid in knowing whether interference affects estimates. Scholars may examine whether units interfere with each other and in which direction the bias is likely to be. Another solution is designing studies with the purpose of mitigating the likelihood of violations. For instance, researchers may cluster randomize the treatment to mitigate the potential for interference to bias conclusions. If researchers can identify a level at

which it is unlikely that units interfere, randomizing the treatment at this level may be the most appropriate option. This is done in several studies on coproduction, in which researchers randomize language support at the level of child care centers rather than the level of families (Jakobsen 2013; Jakobsen and Andersen 2013a; Thomsen and Jakobsen 2015). Though this decreases statistical power, it also mitigates the issue that families may help friends in their vicinity who do not receive language support.

Threat 3: Attrition. Attrition occurs when outcome data are missing. Attrition becomes a problem for causal inference when two conditions are present: (1) units with missing outcomes differ systematically on the outcome from those that are not missing and (2) attrition is different in experimental groups (James, John, and Moseley 2017, 106–107). There is greater potential for attrition in field experiments than in laboratory experiments (James, John, and Moseley 2017). This is partly because field experiments confer less control, making it more difficult to measure the outcome variables. There are generally two sorts of solutions to attrition: design-based solutions and statistical solutions.

Designing and implementing experiments with the goal of minimizing attrition is a solution that researchers should consider (James, John, and Moseley 2017). Jensen (2018) uses measures to minimize attrition. The treatment was conducted in geographic proximity of the respondents, and scholars on the LEAP project sought to identify reasons for attrition beforehand to find ways to decrease it.

Gerber and Green (2012) suggest several design-based and statistical solutions to accommodate attrition. One of their suggested solutions is to contact nonrespondents as frequently as possible with a given level of resources. A related solution is multistage sampling. In multistage sampling, researchers first sample in clusters, such as geographic areas. Then, the second stage can randomly sample individuals within these clusters (European Social Survey 2018, 12).

The statistical solutions they suggest are bounding data by filling in the most plausible data or the data that would be most conservative in relation to any observed significant treatment effects. Such statistical solutions are based on guesses and do not take counterfactual scenarios into account (Pearl and MacKenzie 2018), but they may be desirable when the aim is to demonstrate that differential attrition does not affect causal inferences. We did not find any public administration field experiments using statistical solutions to attrition.

External Validity

In addition to the identified threats to internal validity, there are also threats to external validity, posed by, among other things, noncompliance and randomization bias. Threats to external validity demonstrate the importance of knowing the quantity of interest. Furthermore, the external validity may not allow field experimenters to infer the scalability of the project because of idiosyncrasies of participating units or implementation issues. Put simply: context may matter. We consider noncompliance and randomization bias to be threats to external validity.

Threat 1: Noncompliance. Noncompliance occurs when subjects do not comply with the experimental condition they were assigned to. When noncompliance occurs, researchers have to choose whether they want to examine the effects of the intervention or the effects of the intervention among those who complied with the treatment. Noncompliance is thus important to consider when choosing the quantity of interest and is closely related to external validity. However, noncompliance can also be an issue of internal validity, because it may bias causal inferences depending on how researchers choose to manage it.

One might measure causal effects in the case of noncompliance in two ways: either through the intent to treat effect (ITT) or through the complier average causal effect (CACE) (Athey and Imbens 2017, 114–115). The ITT is the effect of treatment assignment, ignoring any issues with noncompliance among experimental units. This means that the ITT is the overall effect of the intervention. It is also called the program effect. Scholars might be very interested in this quantity, and decision makers may want to know whether an intervention works as intended. This may increase external validity because noncompliance will likely occur in real-world settings.

Jakobsen and Andersen (2013a, 708) consider this issue in their article on coproduction and conclude that the ITT is more relevant in their study since they show interest in "the effect as it occurs in the actual service delivery setting, including all possible implementation issues." This is a valid reason, since the question of interest is whether coproduction increases equity. The ability to provide knowledge about this depends on treatment assignment rather than whether citizens comply.

Yet, even though the ITT measures the program effect and includes noncompliers, its external validity may be limited when generalizing to a new case in which the mechanisms may work differently. In this case, the CACE is recommendable (Athey and Imbens 2017, 117).

The CACE refers to the effects of the treatment among compliers rather than the effects of treatment assignment. This is measured by the CACE, which is the causal effect of treatment assignment for those who complied with their assigned treatment. If the CACE is of interest, the scholar should conduct a two-stage least squares (2SLS) regression (Gerber and Green 2012, 159). The scholar obtains the CACE, which is the ITT divided by the proportion of compliers. Some field experiments aim to know the CACE. Grimmelikhuijsen and Klijn (2015) examine whether judicial transparency increases public trust and randomly distribute encouragement to watch a television series about judges. Some citizens may not receive the encouragement. Since Grimmelikhuijsen and Klijn are interested in the effect of transparency on trust (and not the effect of the encouragement assignment), they recover the CACE.

The CACE also puts restrictions on external validity. It is a local average treatment effect (LATE), since it only recovers the causal effects among those who comply with their treatment assignment (Aronow and Carnegie 2013). Thus, the CACE does not generalize to the entire population, where implementation difficulties may dilute treatment effects.

Threat 2: Randomization Bias. Another threat to external validity is randomization bias (Heckman and Smith 1995). Randomization bias refers to the notion that those signing up for participation in

randomized controlled trials may differ significantly from those who do not. This creates a form of selection bias, the implication being that research on interventions may not generalize to the entire population. For instance, participating units may be those who are most likely to experience positive effects of the treatment, implying that the intervention yields smaller effects when scaled to the entire population (Andersen and Hvidman 2019).

This is relevant for public administration research, since interventions are often carried out with organizations and there may be systematic differences between those who choose to participate and those who do not. Little inquiry examines whether this is the case, which is why we do not know whether public administration field experiments generalize to the entire set of cases of interest. A new project by Andersen and Hvidman (2019) examines differences between participating schools and parents in a reading intervention and whether background differences or differences in implementation explain the effects of the intervention. Overall, the intervention does not show any apparent effects. Yet the researchers have data on the entire population of schools, which allows them to compare participants with nonparticipants in their study. The study allows for interesting findings as to who participates in field experiments and whether implementation issues explain heterogeneous treatment effects. Results show that weak implementation is a more important issue than heterogeneous treatment effects and selection into trials for this type of interventions. Such a study may provide a first suggestion of how randomization bias and implementation issues affect the scalability of public administration field experiments.

The Future of Field Experimentation in Public Administration

This article has reviewed public administration field experiments. There is an increasing trend in field experimentation in public administration (very recent studies include Linos and Riesch 2020; Sanders and Kirkman 2019). The credibility revolution in economics has diffused into public administration, where a focus on identification of causal effects combined with preferences for realistic treatments explains the proliferation of field experiments. Some topics are increasingly studied using field experiments. These include performance information (Andersen and Larsen 2016), leadership (Bellé 2014), coproduction (Thomsen and Jakobsen 2015), and discrimination (Jilke, van Dooren, and Rys 2018).

However, there are still several topics in public administration where field experiments have yet to make their mark. One is innovation in the public sector. Since innovations are organizational interventions, it seems that such studies would be costly, though they would likely also carry large benefits to science and society (e.g., Clingsmith and Shane 2018; Skiera and Nabout 2013). Similarly, administrative burden is a new field of inquiry in which field experiments may yield inferences of high theoretical and practical relevance. By examining how to reduce the psychological, learning, and compliance costs (Moynihan, Herd, and Harvey 2014) carried by citizens when accessing public services, the literature may take a huge leap forward in informing policy makers (a very recent study is Linos, Quan, and Kirkman 2020).

Finally, the effects of networks on individual information levels, attitudes, and change (e.g., Bandiera and Rasul 2006) offers potential. Little research in public administration assesses the importance of networks on individual attitudes and behavior in field experiments, but these may be studied causally in realistic settings to examine how information flows affect individuals in organizations and communities. We hope that researchers seeking to answer such questions will turn to this review to consider costs, practicality, ethics, and validity of their studies, which could enable more field experimentation in public administration in the future.

Another issue that might benefit the potential for more field experiments in public administration is research into methodological issues in field experiments. We suggest that future research consider examining how to make organizations participate in field experiments, how to increase compliance in field experiments, and how and when interference between experimental units is an issue. It is notoriously difficult to engage organizations in field experiments (Eden 2017), since they typically have different interests than those of the researcher (Levitt and List 2009). Research could examine how different information cues affect organizational and individual responses to participating in field experiments. This would help public administration field experiments mitigate randomization bias (Andersen and Hvidman 2019). Second, noncompliance is an issue limiting the generalizability of field experimental claims. If research examined how to make individuals comply with their assigned treatment and possibly also how to avoid attrition, this would benefit future field experiments in public administration. Finally, interference may bias conclusions in field experiments, which is why information on when interference is an issue would help field experimenters to know when it is preferable to cluster randomize treatments and when it is more desirable to do so at the individual level. Related disciplines have modeled spillover effects in voting campaigns (Sinclair, McConnell, and Green 2012) and school enrollment decisions (Bobonis and Finan 2009), but interference may work differently in public organizational contexts, which is why public administration field experiments may benefit from further inquiry into the issue.

We now turn to our suggestions for conducting field experiments in the future. More credible studies are, first, achieved by focusing on preregistration. We found only two preregistered studies in our review (Grimmelikhuijsen et al. 2019; Menger and Stein 2018). Preregistration carries benefits that public administration field experimenters can utilize (Nosek et al. 2018). Preregistration enables scholars to think about design choices before conducting studies. Furthermore, preregistration mitigates the risk of hypothesizing after results are known, a very probable cause of the replication crisis (Open Science Collaboration 2015). Flexibility allows scholars to achieve statistical significance no matter whether the data is indicative of effects (Simmons, Nelson, and Simonsohn 2011). Preregistration allows for scholars to think about the choices they make before gathering data and mitigate "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn 2011), which enhance false positives in research.

Public administration field experiments are becoming increasingly popular, and with good reason. Field experiments may inform about causal effects in realistic settings that other types of studies

seldom inform about. Yet there are few field experiments in public administration. We hope that this systematic review guides researchers in considering how to conduct field experiments from the early phases in ensuring collaboration with organizations or participants, to the design phase where issues of validity can be addressed. By demonstrating how these considerations are discussed in public administration field experiments and providing suggestions on how to find further information, we have sought to provide a primer on field experiments for researchers new to the method. Furthermore, we have suggested future research agendas that we hope field experimenters will attend to in the years to come.

Acknowledgments

We are grateful to Donald P. Moynihan, Morten Jakobsen, and Simon Calmar Andersen, who helped us find articles for this review. Furthermore, we want to thank Peter John and three anonymous reviewers for their helpful comments. Lars Tummers acknowledges funding of NWO Grant 016.VIDI.185.017. Furthermore, he acknowledges that this work was supported by the National Research Foundation of Korea Grant, funded by the Korean Government (NRF-2017S1A3A2067636).

References

Note: All references included in this review are listed here and denoted by an asterisk.

*Adman, Per, and Hanna Jansson. 2017. A Field Experiment on Ethnic Discrimination among Local Swedish Public Officials. *Local Government Studies* 43(1): 44–63.

*An, Seung-Ho, Kenneth J. Meier, Anne Bøllingtoft, and Lotte Bøgh Andersen. 2019. Employee Perceived Effect of Leadership Training: Comparing Public and Private Organizations. *International Public Management Journal* 22(1): 2–28.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

———. 2010. The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives* 24(2): 3–30.

Andersen, Simon Calmar, and Ulrik Hvidman. 2019. Implementing Parent-Aimed Programs at Scale: Evidence from a Randomized Controlled Trial. Working paper.

*Andersen, Simon Calmar, and Morten H. Larsen. 2016. Cognitive Biases in Performance Evaluations. *Journal of Public Administration Research and Theory* 26(4): 647–62.

*Andersen, Simon Calmar, and Donald P. Moynihan. 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): 448–60.

———. 2018. How Do Socially Distinctive Newcomers Fare? Evidence from a Field Experiment. *Public Administration Review* 78(6): 874–82.

*Andersen, Simon Calmar, Helena S. Nielsen, and Mette Kjærgaard Thomsen. 2018. How to Increase Citizen Coproduction: Replication and Extension of Existing Research. *International Public Management Journal*. Published online December 20. https://doi.org/10.1080/10967494.2018.1518851.

Aronow, Peter M., and Allison Carnegie. 2013. Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable. *Political Analysis* 21(4): 492–506.

Athey, Susan, and Guido W. Imbens. 2017. The Econometrics of Randomized Experiments. In *Handbook of Field Experiments*, edited by Abhijit V. Banerjee and Esther Duflo, 73–140. Oxford: Elsevier.

Baekgaard, Martin, Caroline Baethge, Jens Blom-Hansen, Claire A. Dunlop, Marc Esteve, Morten Jakobsen, Brian Kisida, John Marvel, Alice Moseley, Søren Serritzlew, Patrick Stewart, Mette Kjærgaard Thomsen, and Patrick J. Wolf. 2015. Conducting Experiments in Public Management Research: A Practical Guide. *International Public Management Journal* 18(2): 323–42.

Bandiera, Oriana, and Imran Rasul. 2006. Social Networks and Technology Adoption in Northern Mozambique. *Economic Journal* 116(514): 869–902.

Battaglio, Paul R., Jr., Paolo Belardinelli, Nicola Bellé, and Paola Cantarelli. 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): 304–20.

*Bellé, Nicola. 2013. Experimental Evidence on the Relationship between Public Service Motivation and Job Performance. *Public Administration Review* 73(1): 143–53.

———. 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): 109–36.

———. 2015. Performance-Related Pay and the Crowding Out of Motivation in the Public Sector: A Randomized Field Experiment. *Public Administration Review* 75(2): 230–41.

*Bellé, Nicola, and Paola Cantarelli. 2019. Do Ethical Leadership, Visibility, External Regulation, and Prosocial Impact Affect Unethical Behavior? Evidence from a Laboratory and a Field Experiment. *Review of Public Personnel Administration* 39(3): 349–71.

*Bellé, Nicola, Paola Cantarelli, and Paolo Belardinelli. 2017. Cognitive Biases in Performance Appraisal: Experimental Evidence on Anchoring and Halo Effects with Public Sector Managers and Employees. *Review of Public Personnel Administration* 37(3): 275–94.

*Ben-Aaron, James, Matthew Denny, Bruce Desmarais, and Hanna Wallach. 2017. Transparency by Conformity: A Field Experiment Evaluating Openness in Local Governments. *Public Administration Review* 77(1): 68–77.

*Bhanot, Syon P., Gordon Kraft-Todd, David Rand, and Erez Yoeli. 2018. Putting Social Rewards and Identity Salience to the Test: Evidence from a Field Experiment with Teachers in Philadelphia. *Journal of Behavioral Public Administration* 1(1): 1–16.

Bobonis, Gustavo J., and Frederico Finan. 2009. Neighborhood Peer Effects in Secondary School Enrollment Decisions. *Review of Economics and Statistics* 91(4): 695–716.

Bowers, Jake, Mark M. Frederickson, and Costas Panagopoulos. 2013. Reasoning about Interference between Units: A General Framework. *Political Analysis* 21(1): 97–124.

Brady, Henry E., and David Collier. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield.

Broockman, David E., Joshua L. Kalla, and Jasjeet S. Sekhon. 2017. The Design of Field Experiments with Survey Outcomes. A Framework for Selecting More Efficient, Robust, and Ethical Designs. *Political Analysis* 25(4): 435–64.

Clingsmith, David, and Scott Shane. 2018. Training Aspiring Entrepreneurs to Pitch Experienced Investors: Evidence from a Field Experiment in the United States. *Management Science* 64(11): 5164–79.

Czibor, Eszter, David Jimenez-Gomez, and John A. List. 2019. The Dozen Things Experimental Economists Should Do (More of). *Southern Economic Journal* 86(2): 371–432.

*Darolia, Rajeev, Cory Koedel, Paco Martorell, Katie Wilson, and Francisco Perez-Arce. 2015. Do Employers Prefer Workers Who Attend For-Profit Colleges? Evidence from a Field Experiment. *Journal of Policy Analysis and Management* 34(4): 881–903.

*Dudau, Adina, Georgios Kominis, and Melinda Szocs. 2018. Innovation Failure in the Eye of the Beholder: Towards a Theory of Innovation Shaped by Competing Agendas within Higher Education. *Public Management Review* 20(2): 254–72.

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Eden, Dov. 2017. Field Experiments in Organizations. *Annual Review of Organizational Psychology and Organizational Behavior* 4: 91–122.

European Social Survey. 2018. European Social Survey Round 9 Sampling Guidelines: Principles and Implementation. https://www.europeansocialsurvey.org/docs/round9/methods/ESS9_sampling_guidelines.pdf [accessed February 27, 2020].

Falk, Armin, and James J. Heckman. 2009. Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science* 23: 535–8.

*Feigenberg, Benjamin, Erica Field, Rohini Pande, Natalia Rigol, and Shayak Sarkar. 2014. Do Group Dynamics Influence Social Capital Gains among Microfinance Clients? Evidence from a Randomized Experiment in Urban India. *Journal of Policy Analysis and Management* 33(4): 932–49.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton.

Glennerster, Rachel. 2017. The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency. In *Handbook of Field Experiments*, edited by Abhijit V. Banerjee and Esther Duflo, 175–243. Oxford: Elsevier.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations*. Princeton, NJ: Princeton University Press.

Grimmelikhuijsen, Stephan, Sebastian Jilke, Asmus Leth Olsen, and Lars G. Tummers. 2017. Behavioral Public Administration: Combining Insights from Public Administration and Psychology. *Public Administration Review* 77(1): 45–56.

*Grimmelikhuijsen, Stephan, Peter John, Albert Meijer, and Ben Worthy. 2019. Do Freedom of Information Laws Increase the Transparency of Government? A Pre-registered Replication of a Field Experiment. *Journal of Behavioral Public Administration* 1(2): 1–10.

*Grimmelikhuijsen, Stephan, and Albert Klijn. 2015. The Effects of Judicial Transparency on Public Trust: Evidence from a Field Experiment. *Public Administration* 93(4): 995–1011.

*Grohs, Stephan, Christian Adam, and Cristoph Knill. 2016. Are Some Citizens More Equal than Others? Evidence from a Field Experiment. *Public Administration Review* 76(1): 155–64.

*Haynes, Laura C., Donald P. Green, R. Gallagher, Peter John, and David J. Torgerson. 2013. Collection of Delinquent Fines: An Adaptive Randomized Trial to Assess the Effectiveness of Alternative Text Messages. *Journal of Policy Analysis and Management* 32(4): 718–30.

Heckman, James J., and Jeffrey A. Smith. 1995. Assessing the Case for Social Experiments. *Journal of Economic Perspectives* 9(2): 85–110.

*Hess, Douglas R., Michael J. Hanmer, and David W. Nickerson. 2016. Encouraging Local Compliance with Federal Civil Rights Laws: Field Experiments with the National Voter Registration Act. *Public Administration Review* 76(1): 165–74.

*Hock, Scott, Sarah Anderson, and Matthew Potoski. 2013. Invitation Phone Calls Increase Attendance at Civic Meetings: Evidence from a Field Experiment. *Public Administration Review* 73(2): 221–8.

*Jacobsen, Christian Bøtcher, and Lotte Bøgh Andersen. 2019. High Performance Expectations: Concept and Causes. *International Journal of Public Administration* 41(11): 1–11.

*Jakobsen, Morten. 2013. Can Government Initiatives Increase Citizen Co-production? Results of a Randomized Field Experiment. *Journal of Public Administration Research and Theory* 23(1): 27–54.

*Jakobsen, Morten, and Simon Calmar Andersen. 2013a. Coproduction and Equity in Public Service Delivery. *Public Administration Review* 73(5): 704–13.

———. 2013b. Intensifying Social Exchange Relationships in Public Organizations: Evidence from a Randomized Field Experiment. *Journal of Policy Analysis and Management* 32(1): 60–82.

*James, Oliver. 2010. Performance Measures and Democracy: Information Effects on Citizens in Field and Laboratory Experiments. *Journal of Public Administration Research and Theory* 21(3): 399–418.

———. 2011. Managing Citizens' Expectations of Public Service Performance: Evidence from Observation and Experimentation in Local Government. *Public Administration* 89(4): 1419–35.

*James, Oliver, and Alice Moseley. 2014. Does Performance Information about Public Services Affect Citizens' Perceptions, Satisfaction, and Voice Behavior? Field Experiments with Absolute and Relative Performance Information. *Public Administration* 92(2): 493–511.

James, Oliver, Peter John, and Alice Moseley. 2017. Field Experiments in Public Management. In *Experiments in Public Management Research: Challenges and Contributions*, edited by Oliver James, Sebastian R. Jilke, and Gregg G. van Ryzin, 89–116. Cambridge, Cambrige University Press.

*Jensen, Ulrich T. 2018. Does Perceived Societal Impact Moderate the Effect of Transformational Leadership on Value Congruence? Evidence from a Field Experiment. *Public Administration Review* 78(1): 48–57.

*Jilke, Sebastian, Wouter van Dooren, and Sabine Rys. 2018. Discrimination and Administrative Burden in Public Service Markets: Does a Public–Private Difference Exist? *Journal of Public Administration Research and Theory* 28(3): 423–39.

*Jilke, Sebastian, Jiahuan Lu, Chengxin Xu, and Shugo Shinohara. 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.

*John, Peter, and Toby Blume. 2018. How Best to Nudge Taxpayers? The Impact of Message Simplification and Descriptive Social Norms on Payment Rates in a Central London Local Authority. *Journal of Behavioral Public Administration* 1(1). https://doi.org/10.30636/jbpa.11.10.

*Kisida, Brian, and Patrick J. Wolf. 2015. Customer Satisfaction and Educational Outcomes: Experimental Impacts of the Market-Based Delivery of Public Education. *International Public Management Journal* 18(2): 265–85.

Levitt, Steven D., and John A. List. 2007. What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World? *Journal of Economic Perspectives* 21(2): 153–74.

———. 2009. Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review* 53(1): 1–18.

*Linos, Elizabeth. 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.

Linos, Elizabeth, Lisa T. Quan, and Elspeth Kirkman. 2020. Nudging Early Reduces Administrative Burden: Three Field Experiments to Improve Code Enforcement. *Journal of Policy Analysis and Management* 39(1): 243–65.

*Linos, Elizabeth, Joanne Reinhard, and Simon Ruda. 2017. Levelling the Playing Field in Police Recruitment: Evidence from a Field Experiment on Test Performance. *Public Administration* 95: 943–56.

Linos, Elizabeth, and Nefara Riesch. 2020. Thick Red Tape and the Thin Blue Line: A Field Study on Reducing Administrative Burden in Police Recruitment. *Public Administration Review* 80(1): 92–103.

List, John A. 2011. Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off. *Journal of Economic Perspectives* 25(3): 3–16.

Lonati, Sirio, Bernardo F. Quiroga, Christian Zehnder, and John Antonakis. 2018. On Doing Relevant and Rigorous Experiments: Review and Recommendations. *Journal of Operations Management* 64(1): 19–40.

*Menger, Andrew, and Robert M. Stein. 2018. Enlisting the Public in Facilitating Election Administration: A Field Experiment. *Public Administration Review* 78(6): 892–903.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine* 151(4): 264–9.

Moynihan, Donald P., Pamela Herd, and Hope Harvey. 2014. Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions. *Journal of Public Administration Research and Theory* 25(1): 43–69.

Mutz, Diana. 2011. *Population-Based Survey Experiments*. Princeton: Princeton University Press.

Nosek, Brian, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The Preregistration Revolution. *Proceedings of the National Academy of Sciences* 115(11): 2600–6.

Open Science Collaboration. 2015. Estimating the Reproducibility of Psychological Science. *Science* 349(6251): aac4716.

Pager, Devah. 2007. The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future. *Annals of the American Academy of Political and Social Science* 609: 104–33.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

*Peisakhin, Leonid, and Paul Pinto. 2010. Is Transparency an Effective Anti-corruption Strategy? Evidence from a Field Experiment in India. *Regulation & Governance* 4(3): 261–80.

Pfaff, Steven, Charles Crabtree, Holger Kern, and John B. Holbein. 2019. Does Religious Bias Shape Access to Public Services? A Large-Scale Audit Experiment among Street-Level Bureaucrats. SocArXiv. https://osf.io/preprints/socarxiv/9khds/.

*Porter, Ethan, and John C. Rogowski. 2018. Partisanship, Bureaucratic Responsiveness, and Election Administration: Evidence from a Field Experiment. *Journal of Public Administration Research and Theory* 28(4): 602–17.

Sanders, Michael, and Elspeth Kirkman. 2019. I've Booked You a Place, Good Luck: A Field Experiment Applying Behavioral Science to Improve Attendance at High Impact Recruitment Events. *Journal of Behavioral Public Administration* 2(1). https://doi.org/10.30636/jbpa.21.24.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* 22(11): 1359–66.

Sinclair, Betsy, Margaret McConnell, and Donald P. Green. 2012. Detecting Spillover Effects: Design and Analysis of Multilevel Experiments. *American Journal of Political Science* 56(4): 1055–69.

Skiera, Bernd, and Abou Nabout. 2013. PROSAD: A Bidding Decision Support System for Profit Optimizing Search Engine Advertising. *Marketing Science* 32(2): 213–20.

*Thomsen, Mette Kjærgaard, and Morten Jakobsen. 2015. Influencing Citizen Coproduction by Sending Encouragement and Advice: A Field Experiment. *International Public Management Journal* 18(2): 286–303.

*Vashdi, Dana R. 2013. Teams in Public Administration: A Field Study of Team Feedback and Effectiveness in the Israeli Public Healthcare System. *International Public Management Journal* 16(2): 275–306.

*Worthy, Ben, Peter John, and Matia Vannoni. 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.

Supporting Information

A supplementary appendix may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1111/puar.13181/full.