Received 6 July 2015
Revised 15 November 2015
Accepted 15 December 2015

Experimental public administration from 1992 to 2014

A systematic literature review and ways forward

Robin Bouwman
Public Administration Department, Institute for Management Research, Radboud University Nijmegen, Nijmegen, The Netherlands, and
Stephan Grimmelikhuijsen
Utrecht School of Governance, Utrecht University, Utrecht, The Netherlands

Abstract

Purpose – Based on previous inventories, the purpose of this paper is to extend the knowledge on public administration experiments by focusing on their experimental type, design, sample type and realism levels and external validity. The aim is to provide an overview of experimental public administration and formulate potential ways forward.

Design/methodology/approach – The authors examine the current state of experimental public administration, by looking at a systematic selection of ISI ranked experimental publications in major public administration journals (1992-2014) and recommend ways forward based on this review.

Findings – The review indicates a rise in experimentation in public administration in recent years, this can be attributed mostly to some subfields of public administration. Furthermore, most experiments in public administration tend to have relatively simple designs, high experimental realism and a focus on external validity. Experimental public administration can be strengthened by increasing diversification in terms of samples, experimental designs, experimental types and substantive scope. Finally, the authors recommend to better utilize experiments to generate usable knowledge for practitioners and to replicate experiments to improve scientific rigour.

Originality/value – This paper contributes to experimental public administration by drawing on a systematic selection of papers and assessing them in depth. By means of a transparent and systematic selection of publications, various venues or ways forward are presented.

Keywords Causality, Experiments, Literature review, Methodological innovation, Public management research, Replication

Paper type Literature review

1. Introduction

In contrast with other social sciences, such as psychology, economy and more recently political science, public administration research does not have a tradition of experimentation. Various leading public administration scholars have called for more experimentation in our field in addition to the current methodological toolkit

An earlier version of this paper was presented at the IRSPM conference 2014 in Ottawa, Canada. The authors want to thank two anonymous reviewers and the editor-in-chief, Sandra van Thiel, for their constructive feedback on this paper. Also the feedback by Joram Feitsma was helpful in revising the manuscript.

Funding: this work is financed by the Netherlands Organization for Scientific Research (NWO), Research Talent Grant No. 406-13-021.

International Journal of Public Sector Management
Vol. 29 No. 2, 2016
pp. 110-131

0951-3558
DOI 10.1108/IJPSM-07-2015-0129

(e.g. Margetts, 2011; Perry, 2012), which is currently mainly focused at qualitative and survey research (Groeneveld *et al.*, 2015).

Why is experimentation such a useful additional method for public administration? First and foremost, experimental designs provide a solution to the problem of endogeneity that is often present in social science research (Morton and Williams, 2010; Angrist and Pischke, 2008). There are many potential causes of endogeneity. One common cause of endogeneity is the presence of an uncontrolled confounding variable, which influences both the independent and dependent variables of a model. Second is simultaneity, for instance, when there is a correlation between two variables but we do not know the direction of the causal effect. Other causes of endogeneity relate to, for instance, selection effects, measurement bias and common method bias, yet it goes beyond the scope of this paper to discuss them all here (see Podsakoff *et al.*, 2003 for an overview). To overcome endogeneity – particularly omitted variable bias and simultaneity – an experimental design is crucial.

Why is an experimental design crucial to overcome endogeneity? The idea behind an experiment is that the researcher deliberately manipulates one potential variable of interest that could cause an effect in another variable, let's say the effect of representative bureaucracy on perceived legitimacy. In this case, Riccucci *et al.* (2014) manipulated the degree of representativeness in a vignette, holding all other relevant factors constant. Together with proper randomization (explained later in this paper) the problem of simultaneity is avoided and researchers can much easier control for omitted variables.

As a result, experiments have the potential to answer causal questions and contribute to theory development. For instance, experimentation resulted in adjustment of Niskanen's model of budget maximizing behaviour bureaucrats (Moynihan, 2013). Niskanen's (1971) model predicts that bureaucrats will always advocate larger budgets in order to increase their power. However, this premise has been challenged because people working in the public sector are said to also pursue societal interest and not (only) self-interest (i.e. public service motivation (PSM)). Moynihan (2013) tests this alternative explanation by means of a survey experiment. In Moynihans' (2013) experiment, participants make budget decisions based on short hypothetical scenarios. Subjects differed on the degree of PSM and Moynihan (2013) found that individuals with higher PSM are not mere budget maximizers in their decision making (Moynihan, 2013).

There are numerous other examples of how experiments can contribute to answer causal questions in public management, such as the effect of civil servant motivation and organizational performance (e.g. Bellé, 2013), the effect of judicial transparency and citizen trust (Grimmelikhuijsen and Klijn, 2015) and the effects of credibility of performance information on citizen satisfaction (James and Moseley, 2014; Olsen, 2015).

These examples indeed fit neatly with the calls for more experimentation (e.g. Perry, 2012) and some have concluded there is a rise of experimentation in public administration (e.g. Margetts, 2011). In this paper, we contribute to this burgeoning method in public administration research by answering the following question: how have experiments in public administration been designed thus far and – based on this – what are future directions future experimental public administration?

There have been valuable previous inventories on experimental public administration, for instance, by Margetts (2011) and Anderson and Edwards (2015). What does this paper add to this literature? First, we provide a systematic assessment of empirical experimental articles in public administration research in the period from 1992 through 2014. Second, we assess these experiments on some crucial characteristics: design and type of experiment, sample and type of participants, and external validity. In doing so, our

research builds on recent inventories of experiments with very useful recommendations by Margetts (2011) and Anderson and Edwards (2015). In sum, we contribute to this literature by carrying out a detailed and systematic analysis of the characteristics of experiments in our field, which is then used to identify ways forward for experimental public administration.

2. A framework to assess experimental public administration

For our analysis of experiments in public administration, we develop a framework based on the following core elements: experimental type and design, sample and external validity. Each element will be specified below.

2.1 Experimental types and design

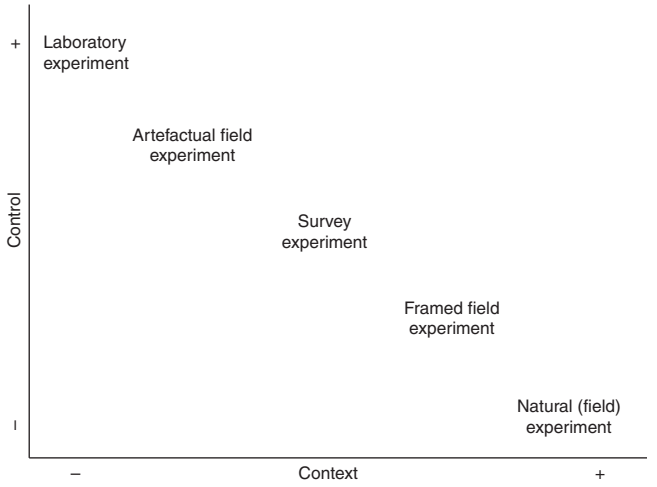
A number of different definitions of experiments exist (see Angrist and Pischke, 2008; Druckman, 2011; Margetts, 2011; Morton and Williams, 2010), but there seems to be agreement on the core elements of experimentation. An experiment has a manipulation, experimental groups of treatment and control, subjects are randomly assigned to experimental groups and there is a systematic method of measuring outcomes.

Based on these elements various types of experiments can be discerned, along two basic dimensions: experimenter control and context. Control is viewed as the extent to which the experimenter has influence over what participants are exposed to during the experiment (Morton and Williams, 2010, p. 305). In a laboratory, a researcher has more control over the effects of the environment on the subjects than in a field experiment.

Experimental context is determined by the nature of the subject pool, the nature of information that subjects bring to the task, the nature of the commodity, the nature of the task or rules of trade, the nature of the stakes and the physical environment in which subjects operate (Harrison and List, 2004, p. 1010). The aforementioned elements of context will oftentimes be associated with each other (Harrison and List, 2004, p. 1012; Morton and Williams, 2010). As a result, the following taxonomy should be considered archetypical (see Harrison and List, 2004, p. 1012).

Harrison and List (2004, p. 1014) identify four types of experiments. A conventional laboratory experiment (uses students as sample, abstract framing and imposed rules), an artefactual field experiment (the same as a laboratory experiment but with a non-student subject-sample like practitioners), a framed field experiment (like the artefactual field experiment but with a field context in task or information set of subjects) and a natural (field) experiment (the same as a framed field experiment but the researcher relies on a natural variation in tasks of subjects and subjects are unaware that they are in an experiment). To this taxonomy, we add the survey experiment. A survey experiment is understood as "An individual decision-making experiment embedded in a survey" (see Morton and Williams, 2010, p. 279). Survey experiments may be paper based, employed over the internet or vignette studies (Morton and Williams, 2010, p. 280) (Figure 1).

Various types of experiments can also be found in public administration research. For instance, Bellé (2013) carried out a framed field experiment with nurses in a hospital in order to investigate the relation between job characteristics and employee performance in the public sector. In the experiment, nurses were exposed to four different treatments, for example a transformational leader, and contact with patients or a self-persuasive brainstorm session (Bellé, 2013, p. 8). One of the findings of this particular experiment is that exposure of employees to transformational leadership in



Source: Adapted from Harrison and List (2004)

Figure 1. Experimental designs

addition to patient contact or a brainstorm session improves performance compared to only exposure to a transformational leader.

Hybrid types can be found too. For instance, (online) survey experiments, in which the manipulation is embedded in a questionnaire, provide less control than a lab experiment. In a survey experiment participants operate in an uncontrolled environment, however, they offer more control than field experiments in terms of manipulation design (e.g. Sulitzeanu-Kenan, 2006; Van de Walle and Van Ryzin, 2011).

Within these types, experiments can have different designs. First, there is the basic design, in which there is one independent variable, a post-test and sometimes a pre-test. The second subtype is the factorial design, which tests the effect of two or more independent variables. For instance, a researcher investigates the effect of exposure to government information about policies on citizen attitudes but is also interested in the effect the content (e.g. negative or positive content). A factorial design then tests all four combinations of both exposure (yes or no) and content (positive or negative). A repeated measurement design takes several measurements sequentially. There are other types of experiments that we will not discuss further because of a lack of space, such as dose-response designs and latin squares designs (see, for more information, Morton and Williams, 2010).

Further, experiments can be between-subjects or within-subjects. In a between-subjects design, the subjects are assigned randomly to either a control or one of the treatment conditions. In a within-subjects design, all subjects receive the control and one or more treatment conditions (Morton and Williams, 2010, pp. 86-87).

In our assessment of experiments in public administration we will gather information on the type of experiment (laboratory experiment, artefactual field experiment, framed field experiment, survey experiment and natural experiment) and design (basic, factorial, repeated measures).

2.2 Composition of the subject-sample

The type of participants used in experiments is an important point of evaluation because it is precisely the subject-sample composition that often is a point of criticism.

Critics point out that experiments too often rely on the use of convenience samples, like student samples. Students are for example not representative for other populations because students are generally highly educated and will therefore react differently to manipulations (Charness and Kuhn, 2011). However, others have argued that student samples are not that problematic and that they even have advantages over other samples. Students samples are more homogeneous which is an advantage, because it is much harder to detect a true effect in heterogeneous samples than it is in homogenous samples (Morton and Williams, 2010, p. 351). Also, some argue that the use of student samples does not truly matter because students do not necessarily react differently to experimental stimuli than other people (Anderson and Edwards, 2015).

We argue that using student samples is not a problem by definition, but should be done with great caution, and results from student samples need further validation in follow-up experiments and replication studies with more heterogeneous samples. Especially if there are reasons to believe that the treatment interacts with some sample characteristics, variation is needed on these characteristics within the sample. For instance, the level of education might influence information processing. Therefore, in a study on the effect of performance information on citizen satisfaction it would make sense to include both lower and higher educated participants in the sample.

In our assessment we assert various sample compositions and we categorize the following: students, practitioners, citizens or mixed samples.

2.3 External validity of experiments

Another point of debate centres around the limited generalizability of experiments (Margetts, 2011; Bozeman and Scott, 1992; Druckman, 2011; Morton and Williams, 2010). This criticism on experiments touches on a broader issue which is their alleged lack of external validity. Indeed, experimental results may not necessarily be generalizable in a statistical sense (i.e. from sample to population). However, the causal effect that is studied is derived from theory and it is the adjustments to theory that can be applied beyond the specific research setting (Calder *et al.*, 1982). So, if we want to assess the external validity of experiments we mean the extent to which it allows for theoretical generalization. This criterion is different than the mere representativeness of the sample.

Shadish *et al.* (2002) distinguish four dimensions of external validity that can be applied to experiments: first, the degree of resemblance between participants and the people who are normally confronted with these stimuli; second, whether the context in which actors operate resembles the context of interest; third, if the stimulus (treatment) resembles a stimulus of interest in the real world; and fourth, whether the outcome measures resemble the actual outcomes of theoretical and practical interest. We decided not to include the fourth criterion in our review, because we found that virtually all studies in public administration had a measure that resemble outcomes of theoretical and practical interest. This may be due to the practical and contextual nature of public administration as a discipline (see Raadschelders and Lee, 2011). Therefore, the first three criteria will be applied in assessing public administration experiments.

The first three dimensions of Shadish *et al.* (2002) directly relate to the debate about the supposed lack of realism of experiments. This has been identified as one of the barriers for using experiments in public administration research (Bozeman and Scott, 1992; Margetts, 2011). More specifically, Bozeman and Scott (1992, p. 309) argue that experiments in public administration research require attention to "mundane realism". Mundane realism is the extent to which the task presented to participants is realistic to them. For instance, this can be improved when the tasks that are carried out in the experiment

are relevant to the tasks that are carried out in reality (Anderson and Edwards, 2015). For instance, if there is no resemblance between participants and people who are normally confronted with the stimulus, in context and with tasks that do not fit what an experiment intends to measure in the real world, their external validity is relatively low. We will assess the external validity of public administration experiments based on these three criteria.

2.4 Framework for assessing experimental public administration

Table I sums the criteria that emerged from the theory. These questions will be used as framework for assessing the experimental practices in public administration research. In addition, we included some more general questions to get a more comprehensive view of experimental public administration.

3. Method and data

We conducted a systematic literature review according to the PRISMA guidelines (Tranfield *et al.*, 2003). First, we selected the public administration category from the ISI-Web of Science database. From this category we listed the top-30 of public administration journals based on five-year impact factor (ISI-Web of Science) (Figure 2).

From this top-30 we selected all public administration journals ($n = 14$)[1], we manually searched on the respective websites of our selected body of journals for papers published from 1992 onwards to 2014. We searched on the journal websites and retained papers that have "Experiment" or "Experimental" in title OR topic OR abstract OR in the keywords ($n = 2{,}521$).

We expect authors to mention that their study employs an experiment at least in the abstract, title or keywords. This search strategy generates a lot of false positives: papers that use the "experiment" to refer to something new (i.e. policy or experimentalist governance).

Next, we evaluated the body of papers on our inclusion criteria. See Table II for all inclusion criteria. We manually filtered out one paper that did meet our criteria but concerns a policy experiment and is focused on economics rather than public administration (i.e. Dynarski *et al.*, 2013). The final body of papers consists of 42 public administration experiments in PA journals.

	Assessment questions
General assessment	In what year was the experiment published?
	In which journal was the experiment published?
	In what kind of substantive field was the experiment published?
Assessing design	What type of experiment is carried out (laboratory experiment, artefactual field experiment, framed field experiment, survey design experiment, natural (field) experiment)
	What is the design of the experiment? (basic; factorial; repeated measures design)
	Is the experiment a within- or between-subjects design?
Assessing samples	What is the sample size of the experiment?
	What types of participants are used in public administration experiments?
Assessing external validity	Do the subjects have resemblance to the people who are normally confronted with these stimuli?
	Does the context in which actors operate resemble the context of interest?
	Does the stimulus (treatment) resemble a stimulus of interest in the real world?

Table I. Framework for assessing public administration experiments

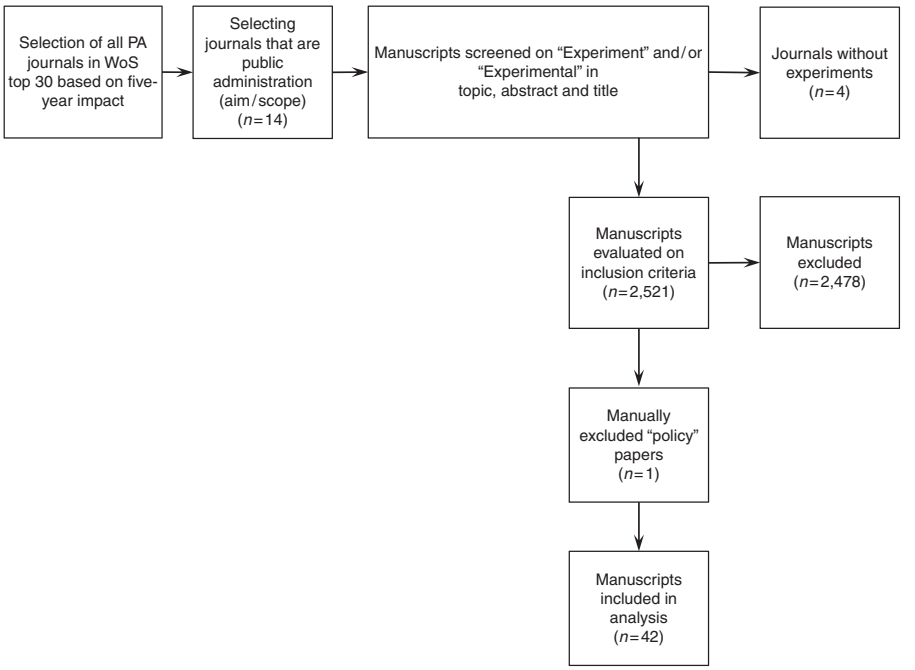


Figure 2.
PRISMA flow diagram

Field	Public administration
Method	Studies should employ an experimental design. Based on our definition this entails that there is an intervention or manipulation, subjects are randomly assigned to groups and there is a systematic method of observing outcomes
Study design	Only empirical studies were included
Type of publication	Only ISI rated international peer-reviewed articles were included
Year of publication	The publication of Bozeman and Scotts' (1992) article in 1992 marks the onset of our selection, 2014 the end (1992; 2014)
Language	English

Table II.
Inclusion criteria

After making a definitive selection, the articles were coded using the operationalization presented in Table AI. Because the items to assess external validity are to some extent subjective, we sought to establish inter-subjectivity on these items. The coding was done in sequential steps. First, both authors analysed the entire collection of publications separately. Second, these separate codings were discussed and merged into a joint coding protocol. Finally, the entire data set was recoded using the refined coding scheme that resulted from the first round of coding. The following Cohen's κ's were retained after the final round of coding: "the subjects have resemblance to the people who are normally confronted with these stimuli" (Cohen's $\kappa = 0.96$) ($p < 0.001$), 95 per cent CI (0.87, 1.04); "Whether the context in which actors operate resembles the context of interest" (Cohen's κ 0.81) ($p < 0.001$), 95 per cent CI (0.72, 0.89); "Whether the stimulus (treatment) resembles a stimulus of interest in the real world" (Cohen's $\kappa = 1.0$) ($p < 0.001$), 95 per cent

CI (–). According Landis and Koch (1977) $K = 0.40$-0.59 is moderate inter-rater reliability, 0.60-0.79 substantial, and 0.80 outstanding, which means our codings have moderate to outstanding consistency.

4. Assessing experiments in public administration

4.1 General assessment of experiments in ISI ranked PA journals

To answer this question we looked at which journals published experimental studies. Table III shows that almost half of all experimental studies have been published in *Journal of Public Administration and Research Theory* (*JPART*). *Public Administration* (*PA*) and *Public Administration Review* (*PAR*) and *Journal of Policy Analysis and Management* (*JPAM*) share a second place and each have published 12 per cent of our selected body of experiments. Other journals published one or two articles with experiments over the years.

Our selection shows a total of 42 publications over a period of 23 years (1992, 2014). Over this period there are on average 1.82 experimental publications per annum (SD 3.74). Kendall's Tau-c was used in order to statistically test whether the number of publications that employ an experimental design in the field of public administration are indeed rising over time. The result indicates that there is a small statistically significant association between the number of reported experimental studies in the field of public administration and time in years ($\tau = 0.35$, $p < 0.05$) (see Figure 3).

This rise can be attributed mainly to 2013 and 2014. In 2014 seven articles have been published and the low number contrasts sharply with 2013. Because 2015 was not finished when collecting the data and writing the paper this year was not included in the analysis. However, given the two themed issues about experimental public administration that were published in 2015 (Blom-Hansen *et al.*, 2015 in *IPMJ* and James *et al.*, 2016 in *PAR*) this rise is likely to continue.

Also, some journals seem to attract more experiments than others; most experiments are published in *JPART*, *PAR*, *PA* and *JPAM*. The sudden increase in experimentation shows that there is certainly a potential application for experiments in our field.

Although the number of experiments in mainstream public administration journals is on the rise, they still only have a minor share in the total number of empirical articles in these journals. A recent inventory by Groeneveld *et al.* (2015) indicates that quantitative studies are on a rise but experiments are only a fraction of total work published within the field of public administration.

Journal	Count	%
Journal of Public Administration and Research Theory	19	45
Public Administration	5	12
Public Administration Review	5	12
Journal of Policy Analysis and Management	5	12
International Public Management Journal	3	7
Public Management Review	1	2
Public Performance and Management Review	1	2
International Review of Administrative Sciences	1	2
Review of Public Personnel Administration	1	2
Governance	1	2
Total	42	100

Note: Sorted in descending count order

Table III. Number and share of total experimental research articles per journal (1992-2014)

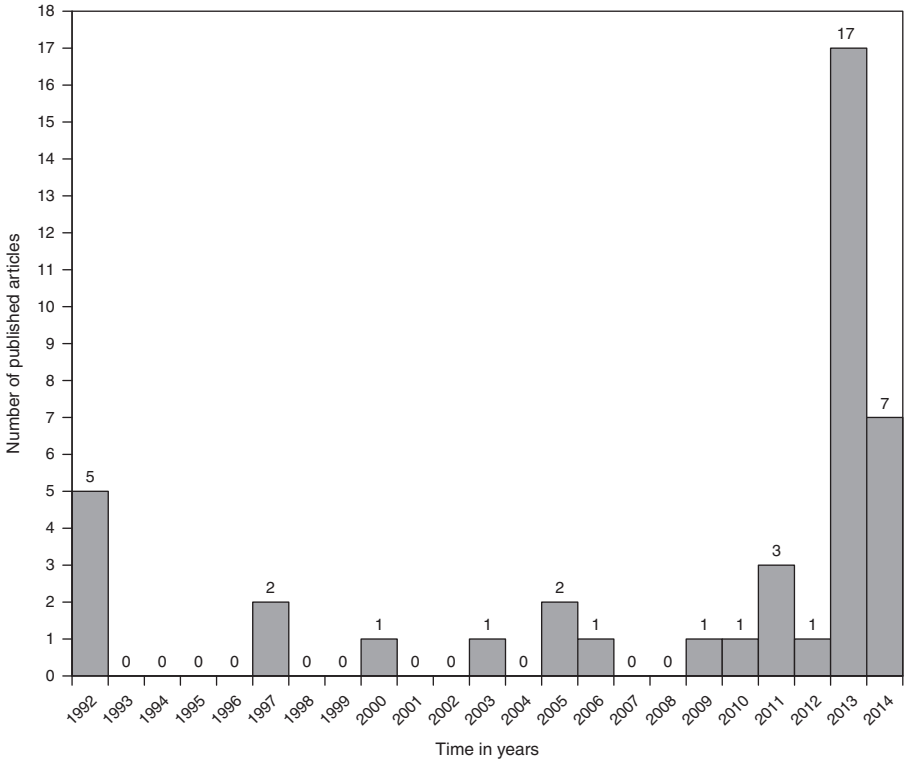


Figure 3.
Published experiments each year

Our assessment also shows a great variety of themes that are investigated by experiments. The studies in our literature selection focus, for example on government performance and decision making, government communication, co-production and transparency (see Table IV). Although research themes addressed by experiments are diverse, there seem to be three main lines of research are present more often than others in our literature selection: first, government performance (e.g. Walker *et al.*, 2013;

Topic	Count	%
Government performance	9	21
Decision making by public servants or citizens	9	21
Government information and communication (not transparency)	4	10
Co-production	3	7
Transparency	3	7
Trust	2	5
Red tape	2	5
Behaviour of individuals (not decision making)	2	5
Motivation and public service motivation	2	5
Others, less than 2 papers (include transparency, innovation, information disclosure, trust, social exchange theory, red tape and satisfaction)	6	14
Total	42	100

Table IV.
Research topics in experimental papers

Bellé, 2013); second, decision making by either citizens or civil servants (e.g. Nielsen and Baekgaard, 2015; Wittmer, 1992); and third, governmental information and communication (e.g. Scott, 1997; Nutt, 2005; James, 2010; Van Ryzin, 2013). We conclude that experiments are not mainstream but rather are more common some subfields of public administration.

4.2 Assessing types and designs of experiments

To assess the design of experiments in public administration, we assessed the method sections of the articles. We looked at the type of experiment (laboratory experiment, artefactual field experiment, framed field experiment, survey experiment and natural experiment), whether the experiment was between or within-subjects and at the specific design (basic; factorial; repeated measures). We plotted our results in Figure 4, which can be found below.

Survey experiments are reported the most; we found fifteen (36 per cent) of them. Second, there were ten (24 per cent) framed field experiments (participants that act in field context). Our selection also shows nine lab experiments (21 per cent). There are two artefactual field experiments (5 per cent), i.e. laboratory experiments but with non-student samples. We found one natural field experiment (2 per cent), in which a natural variation in tasks of subjects and subjects occurs as an experimental "treatment". In five cases (12 per cent), the type of experiment could not be derived from the information in the papers because authors provided too little information or the paper would fit in more than one category because they reported on more than one experiment[2][3].

Most papers reported on a between-subjects design (38 times or 90 per cent). Two studies reported on a within-subjects design (5 per cent). In two cases (5 per cent) this could not be established with certainty because there was too little information in the paper (5 per cent). The between-subjects design seems to be the default option for experimental public administration. One example of an interesting within-subjects design is the study by Christensen *et al.* (2013). The aim of the experiment was to test if PSM affects appraisal outcomes of employees. In this paper-based experiment, MBA and MPA students had to rate employees on their (fictive) performance based on short descriptions (vignettes). All participants received all vignettes; such a within-subjects

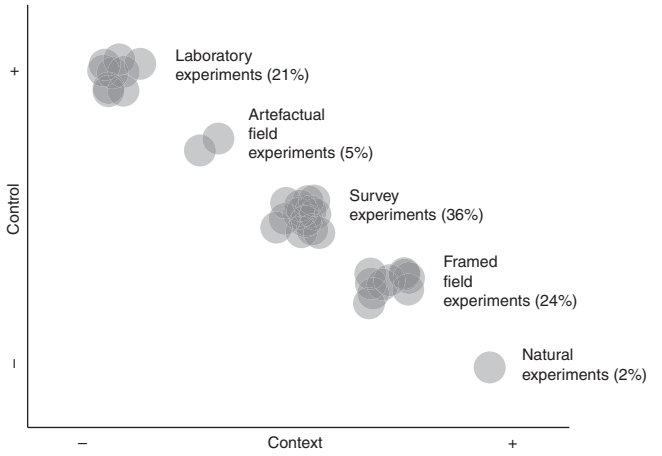


Figure 4.
Experimental control and context plot

design can be a cost-effective way to test a larger number of treatments at the same time, without losing statistical power or having to increase the number of participants.

A closer look at the design of experiments in public administration reveals that a basic design is most commonly applied: this was used in 28 articles (67 per cent) followed by a factorial design (nine times, 21 per cent). Three studies employ a repeated measures design (7 per cent). In one case this is a 3×3 factorial design with repeated measures (2 per cent) For one paper (2 per cent), we could not determine the design because of lacking information. Hence, a basic design is most widely used in experimental public administration. This may be an indication that either researchers are not aware of the existence of alternatives to the basic design or that most research questions in public administration demand a straightforward basic design instead of more challenging alternatives.

4.3 Assessing samples

We took two steps to analyse subject samples, first we looked at overall samples sizes and second we assessed the sample composition.

Table V shows sample size of the experimental types in our literature review.

Table V shows that the overall sample sizes are moderate to large, averaging 470.5 (SD 525.7). This is partly due to a number of large field experiments and one larger natural (field) experiment, which tend to have more participants than laboratory experiments. Survey and internet experiments fall in between in terms of sample size. Furthermore, Table V highlights a high standard deviation in all categories, which indicates that differences among sample sizes of experiments are large. A potential explanation for the large sample sizes in experimental public administration may be that researchers expect small or unknown effect sizes and that therefore a great number of participants per experiment are needed to detect potentially small effects.

We also looked at the composition of the samples and found three types of sampling that are used in public administration experiments. First there are student-only samples and combined student samples. A fairly large proportion of the experiments (16 in total, 38 per cent) in our literature selection mention the use of – graduate and undergraduate – students in their experiments. Four (10 per cent) use both students and a second group such as practitioners. Second, 13 (31 per cent) experiments can be identified that use more or less representative samples of citizens or the general public. Sulitzeanu-Kenan (2006), for instance, uses UK citizens in an online experiment into the effects of public enquiries and the subsequent report on the attributions of public responsibility in the eyes of citizens. Third, there are eight (19 per cent) publications in our literature selection report the use of practitioners like managers or decision makers. In conclusion, the typical public administration experiment employs a variety of samples, but a student sample forms a large proportion of this. Citizens are used mostly

Table V. Sample sizes found in public administration experiments

	Count	Mean	Median	SD
Laboratory Experiment	9	205.3	132.0	203.1
Artefactual field experiment	2	634.5	634.5	562.1
Framed field experiment	10	388.9	319.5	254.4
Survey experiment	15	588.5	400.0	479.4
Natural (field) experiment	1	2,704	–	–
Overall		470.5	280.5	525.7

in online experiments when the research question is aimed at citizens in general. Also we found that various types of samples are used and that samples tend to be fairly large on average. In addition, most scholars seem to use rather simple experimental designs: most employ a basic design, that is, on control and one treatment group.

4.4 Assessing external validity

We used three dimensions to assess the external validity of experiment, which was based on the work of Shadish *et al.* (2002). The result of this assessment is presented in Table VI.

In total, 24 (57 per cent) experiments use samples with a certain degree of resemblance between-subjects and the people who are normally confronted with the experimental stimulus. For instance, these are experiments in which practitioners are used to investigate a research question geared towards practitioners (public managers, policy makers), or student samples when the research is aimed to investigate an effect on the citizen population. A number of authors use both students and non-students to counter potential biases (e.g. Nutt, 2005). In 11 cases (26 per cent) the subjects used have little to no resemblance to the people who are normally confronted with these stimuli. For instance, students are told to play a role that they are not familiar with. In three cases (7 per cent), a combined sample of students and practitioners was used (see section on samples). In these cases, the students may not have resemblance with the people normally confronted with the stimuli while the practitioners in the same experiment do have resemblance. The link between the experimental subjects and target population could not been established in four cases (10 per cent).

The second criterion of Shadish *et al.* (2002) is "whether the context in which actors operate resembles the context of interest", we operationalized this as the extent of "experimental". We admit this is a criterion that is difficult to capture objectively. Therefore, we decided to assess two particular elements of experimental realism that were rather convenient to assess objectively: location and role-playing This may not fully tap into what experimental realism is, but serves as a proxy for it. Results show that ten (24 per cent) experiments were held in a university classroom (including two

Criterion	Findings
Do subjects have resemblance to the people who are normally confronted with these stimuli?	Yes: 57% Somewhat: 7% No: 26% Undetermined: 10%
Does the context in which actors operates resemble the context of interest?	Location University classroom (24%) Computer laboratory (10%) Internet based (33%) Field (24%) Undetermined (10%) Role-play No (83%) Yes (17%)
Does the stimulus (treatment) resemble a stimulus of interest in the real world?	Yes: 88% No: 5% Undetermined: 7%

Table VI.
External validity of experimental designs

vignette studies), four (10 per cent) in a computer laboratory, 14 (33 per cent) experiments involve the use of internet and therefore subjects are involved in them at home, the workplace or elsewhere where an internet connection is available. Ten (24 per cent) experiments were conducted in the field. Some authors do not mention the location and setting of their experiment explicitly (four times, 10 per cent).

The second element of experimental realism concerns whether participants have to play a more or less artificial role or not, e.g. students who have to play the role of an experienced public manager in an experiment. In the selected studies, 35 (83 per cent) experiments do not require subjects to take on a specific role in the experiment. In seven cases (17 per cent) scholars ask subjects to make decisions or perform other actions pretending to be someone else or some official.

The third criterion asserts whether the stimulus resembles a stimulus of interest in the real world. In most public administration experiments in our literature selection, the stimulus is one that has a high resemblance to the "real world" (88 per cent). For example, Scott and Pandey (2000) investigated the effect of red tape on bureaucratic behaviour. In the experiment, levels of red tape were manipulated while the participants reviewed cases for hypothetical clients seeking public assistance; cases that clearly resemble cases that could occur in the daily work of bureaucrats. In three cases it is not quite clear to what extent a treatment has resemblance to the real world (7 per cent) (e.g. Brewer and Brewer, 2011). In two cases (5 per cent), there is no resemblance to a stimulus of interest the real world as this is not the aim of the experiment. This means that the stimulus was somewhat abstracted from how participants could encounter it in their normal environment. An example of this is the order of questions in a survey on satisfaction with public services, which has a research purpose and does therefore not aim to reflect some "real world" stimulus (see Van De Walle and Van Ryzin, 2011).

In sum, the assessment of external validity of public administration experiments show that scholars pay much attention to experimental realism and external validity of their designs. The implications of these future directions for experimental public administration are explored in the next section.

5. Potential ways forward for experimental public administration

Based on our assessment we present the following main findings. First, there has indeed been a recent rise of published experiments in the major public administration journals. Particularly, *JPART*, *IPMJ*, *JPAM*, *PAR* and *PA* have published a variety of experiments. Second, we have seen that experimental public administration has a strong emphasis on the external validity of experimental designs. Third, in a substantial part of the experiments the sampling is still rather narrow, with a large proportion of studies focusing on convenience and/or student samples. Fourth, we have found that the designs currently used by public administration scholars can be mostly qualified as "basic designs" (i.e. one control, one treatment group). The fifth finding is that various domains in public administration have not yet seen any experimental work, but could benefit from this. Based on these findings we have formulated six potential ways forward for experimental public administration, which are summarized in Table VII.

Our first recommendation is to broaden the substantive scope of experiments. We have identified subfields of public administration in which experimentation is already happening: government performance, decision making by citizens or civil servants and government information. This means that there are many substantive areas that do not yet use experiments. Which (other) areas of research can benefit from

Recommendation	Why?
1. Broaden substantive scope	There is a high concentration of experiments around few topics, other areas could benefit from an experimental approach
2. Broaden sampling	We found a rather large proportion of student samples, more diverse samples can be drawn and are needed for most research questions in public administration
3. Explore other experimental designs	Large number of basic designs was found. In order to prevent a narrow focus, other designs should be explored
4. Explore other experimental types	Some types of experiments are hardly used. Experimental PA could be enriched with more artefactual field experiments and natural field experiments
5. Use experiments for usable knowledge	Experiments can be used to rigorously test administrative practices. Currently this only happens to limited extent
6. Replicate experiments to build credible and coherent body of knowledge	The rise of experimental public administration could lead to a base of credible and coherent evidence, but replication across various research designs and contexts is needed

Table VII. Potential ways forward for experimental public administration

experimentation? In principle, experiments can be applied to many situations in which individuals are the objects of analysis (Anderson and Edwards, 2015). Hence, this can regard citizen attitudes, administrative decision making, civil servant perceptions and/or motivations.

But experiments can be broadened to topics that concern the group level too. Social psychology has provided various insights and experiments of how individuals behave in (small) groups. For instance, the concept of groupthink has helped to explain why sometimes elite political decision makers ignore relevant evidence because of an extreme drive to reach group consensus (Janis 1972; 't Hart, 1990). Since much of administrative and political decision making takes place in small groups, these theories and experiments can be applied too. Experimentation can thus be broadened to various subdomains of public administration, as long as there are causal questions involved and when the level of analysis is at the individual level and/or small group level. Some examples include domains of research that deal with administrative decision making in teams, or public-private negotiations can benefit from experimental designs. The same goes for PSM of public employees: certain policies may be tested in randomized trials to assess which ones influence PSM and under which circumstance they work best (see, for instance, Bellé, 2013).

Our second recommendation is to broaden our opportunities to collect diverse samples. Although many public administration experiments generally have relatively large sample sizes, the use of only student samples (found in 31 per cent of the experiments) would fall short for many important research questions in public administration. It might be a useful first step in establishing tentative ideas about certain relationships, but more complex samples are needed to see if these assumptions hold in other samples too. Perfectly representative experimental groups are often unfeasible and are not necessary to achieve the goal of theoretical generalization. Therefore a samples consisting of a heterogeneous group of participants is an interesting way to see if findings from student samples hold. For instance, experiments could greatly contribute to credible knowledge if they report on two experimental tests at the same time: one study to tentatively test an assumption on a homogeneous student sample and one study to test this in a very different sample.

To reach different samples online (survey) experiments are promising. Online experiments offer a cheap way to experiment with subjects in an environment subjects are accustomed to. Also, many public organizations have their own citizen panels which could give access to a broad range of interested citizens. Again, this is not representative, but it is a way to reach broader and more diverse samples. An interesting venue for experimentation is, for instance, Mechanical Turk (MTurk), which is a cheap and convenient way to reach out to a diverse set of participants. MTurk is an online labour market platform where respondents are paid for small tasks, such as participation in surveys and online experiments. Samples obtained via MTurk are not random, but this is not very problematic given that experiments do not necessarily have to rely on random samples – as opposed to random attribution of respondents to experimental conditions. The quality of MTurk data has been criticized, for instance, participants fill out multiple questionnaires each week and doing so very fast without paying attention to the questions. Although there is empirical evidence that MTurk samples do not yield fundamentally different results (Berinsky *et al.*, 2012; Clifford *et al.*, 2015) researchers should take these disadvantages into account by countering the risk of random questionnaire responses. Also scholars could increase robustness of their experiment by using multiple samples in one study. For instance, one could do a test on a student sample and then on a more diverse MTurk sample. To this date there are only a few public administration studies that have used MTurk samples, but see Jilke *et al.* (forthcoming, 2016) and Marvel (2015) for two recent examples in public administration.

Our third recommendation is to explore the possibilities of more diverse experimental designs. Our results show that public administration scholars often prefer basic designs to other design types. Although simple setups are often elegant, public administration scholars could explore designs that allow testing for more variables at one time, such as factorial designs, or more measurements within one subject (repeated measures or within-subjects design). A factorial design consists of two or more independent variables that are deliberately manipulated. For instance, to investigate the influence of performance information on citizen attitudes it might be useful to manipulate both the design of the information, (e.g. relative or absolute performance) and the actual performance (high or low). A factorial design incorporates multiple independent variables to capture the complexity of a phenomenon. Of course, the main reason for choosing an experimental design should be that it fits the research question at hand, but a too narrow focus on basic design may limit the types of research questions we focus on. Exploring more diverse experimental designs can help to prevent this.

The fourth recommendation is to broaden our scope on experimental types. Our review shows that there many employ survey experiments (36 per cent), framed field experiments and – to a lesser extent – lab experiments (see Figure 4). There is room to extend experimental PA to artefactual field experiments (i.e. getting practitioners into the lab) and natural field experiments (i.e. naive participants in field experiments).

Why is this important? The limitations of certain experimental types could be bypassed by replicating an experiment in a richer context. So, for instance, employing a variety of experimental types could help to test if findings from laboratory experiments hold in less stylized and controlled experimental types, such as artefactual or framed field experiments. Public administration scholars may want to consider experiments that have higher experimental control at first (located on the upper left side of Figure 1) and thus more powerfully demonstrate the effect of one or

two isolated variables. One way to do this is by employing computerized experiments and games in a lab. In our literature selection, only one paper makes use of this approach (see Brewer and Brewer, 2011)[4]. We are not arguing that this is the way all experiments in public administration should be conducted, but it could be part of empirical cycle in which an assumed relation is tested in various experimental settings. For instance, after a relation between two variables has been established, other – less controlled – types can be employed to see if the relation holds in experiments with higher realism.

Our fifth recommendation is using experimentation to generate usable knowledge. This aligns with recent calls for experimentation by various scholars (Perry, 2012; James *et al.*, 2016). An interesting example from the private sector is Google's HR department (called People Operations). People Operations continuously experiments with policies, bonuses and other incentives to keep workers happy and motivated (Manjoo, 2013). Obviously the overarching aim is to see how workers are productive and creative. That Google is able to run randomized trials with their employees may pave the way for governments to do the same with their public officials. It seems that not the subject of interest is most relevant for the design, but rather whether the question involves a causal mechanism. Although there are some ethical obstacles to overcome, governments –with the help of social scientists – may be able to experiment with policies when these involve citizens and public employees.

Finally, it is recommended to design and replicate experiments to build a coherent and credible body of knowledge. Because experimental public administration is still in an infant stage, we think that there is great potential and a need to replicate experiments. One of the aims of replication is testing the robustness of results and to apply known results to new settings or subjects. Therefore, replication is a basic requirement for scientific integrity (Open Science Collaboration, 2015). Experiments offer potential for replication as they are more easily replicated than other less structured methods.

The importance of replication can be illustrated by the recent reproducibility study in psychology. A collaboration of over hundred psychological scholars replicated 100 key experiments in their field to assert if the main findings could be replicated. A large portion of replications produced weaker evidence than the original experiments (Open Science Collaboration, 2015) and stresses the importance of replication. In public administration replication can be used to assess if evidence holds across various contexts and research designs. It is crucial that the rise of experimental public administration does not lead to an array of "one shot" experiments, but on building a base of credible evidence that has been tested and replication in various research designs and contexts.

6. Conclusion

Our review shows there has been a remarkable rise of reported experiments in public administration journals in recent years and in leading journals. Also, experiments tend to have high experimental realism and use rather large samples. This trend is likely to continue with two themed issues in public administration journals in 2015 and 2016 (James *et al.*, 2016; Blom-Hansen *et al.*, 2015), a projected handbook of experimental public administration, and a variety of articles that have been published already in 2015 (e.g. Jilke *et al.*, forthcoming, 2016; Marvel, 2015; Tummers *et al.*, 2015). Our study indicates that experimental public administration is on the rise, here to stay, and that there are various ways to go forward and improve experimentation in our field.

Notes

1. In order of five-year impact factor: *Journal of Public Administration and Research Theory*; *Journal of Policy Analysis and Management*, *Governance*; *Public Administration*; *Public Administration Review*; *International Public Management Journal*; *Public Management Review*; *American Review of Public Administration*; *Review of Public Personnel Administration*; *International Review of Administrative Sciences*; *Local Government Studies*; *Administration & Society*; *Public Performance and Management Review* and *Public Money and Management Journal*. Public Policy journals were excluded from the selection.
2. Thurmaier (1992), Landsbergen *et al.* (1992), Scott (1997), Scott and Pandey (2000) and James (2010).
3. Harrison and List (2004, p. 1014) also acknowledge that some studies will not fit neatly into the categories.
4. See, Esteve *et al.* (2015) or Tepe (2015) for more recent examples that are not in our literature selection.

References

*Papers in review.

Anderson, D.M. and Edwards, B.C. (2015), "Unfulfilled promise: laboratory experiments in public management research", *Public Management Review*, Vol. 17 No. 10, pp. 1518-1542.

Angrist, J.D. and Pischke, J.S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton.

*Bellé, N. (2013), "Experimental evidence on the relationship between public service motivation and job performance", *Public Administration Review*, Vol. 73 No. 1, pp. 143-153. doi: 10.111/j.1540-6210.2012.02621.x.Experimental.

Berinsky, A.J., Huber, G.A. and Lenz, G.S. (2012), "Evaluating online labor markets for experimental research: Amazon.com's mechanical turk", *Political Analysis*, Vol. 20 No. 3, pp. 351-368.

Blom-Hansen, J., Morton, R. and Serritzlew, S. (2015), "Experiments in public management research", *International Public Management Journal*, Vol. 18 No. 2, pp. 151-170.

Bozeman, B. and Scott, P. (1992), "Laboratory experiments in public policy and management", *Journal of Public Administration and Research Theory*, Vol. 2 No. 3, pp. 293-313.

*Brewer, G.A. and Brewer, G.A. (2011), "Parsing public/private differences in work motivation and performance: an experimental study", *Journal of Public Administration Research and Theory*, Vol. 21 No. S3, pp. i347-i362.

Calder, B.J., Phillips, L.W. and Tybout, A.M. (1982), "The concept of external validity", *Journal of Consumer Research*, Vol. 9 No. 3, pp. 240-244.

Charness, G. and Kuhn, P.J. (2011), "Lab labor: what can labor economists learn from the lab?", National Bureau of Economic Research Working Paper Series, No. 15913. doi: 10.3386/w15913.

*Christensen, R.K., Whiting, S.W., Im, T., Rho, E., Stritch, J.M. and Park, J. (2013), "Public service motivation, task, and non-task behavior: a performance appraisal experiment with Korean MPA and MBA students", *International Public Management Journal*, Vol. 16 No. 1, pp. 28-52. doi: 10.1080/10967494.2013.796257.

Clifford, S., Jewell, R.M. and Waggoner, P.D. (2015), "Are samples drawn from mechanical turk valid for research on political ideology?", *Research & Politics*, Vol. 2 No. 4. doi: 10.1177/2053168015622072.

Druckman, J. (2011), *Cambridge Handbook of Experimental Political Science*, Cambridge University Press, New York, NY.

Dynarski, S., Hyman, J. and Schanzenbach, D.W. (2013), "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion", *Journal of Policy Analysis and Management*, Vol. 32 No. 4, pp. 692-717.

Esteve, M., van Witteloostuijn, A. and Boyne, G. (2015), "The Effects of public service motivation on collaborative behavior: evidence from three experimental games", *International Public Management Journal*, Vol. 18 No. 2, pp. 171-189.

Groeneveld, S., Tummers, L., Bronkhorst, B., Ashikali, T. and Van Thiel, S. (2015), "Quantitative methods in public administration: their use and development through time", *International Public Management Journal*, Vol. 18 No. 1, pp. 61-86.

Grimmelikhuijsen, S. and Klijn, A. (2015), "The Effects of judicial transparency on public trust: evidence from a field experiment", *Public Administration*, Vol. 93 No. 4, pp. 995-1011.

Harrison, G.W.G. and List, J.A.J. (2004), "Field experiments", *Journal of Economic Literature*, Vol. 42 No. 4, pp. 1009-1055.

*James, O. (2010), "Performance measures and democracy: information effects on citizens in field and laboratory experiments", *Journal of Public Administration Research and Theory*, Vol. 21 No. 3, pp. 399-418.

*James, O. and Moseley, A. (2014), "Does performance information about public services affect citizens' perceptions, satisfaction, and voice behaviour? Field experiments with absolute and relative performance information", *Public Administration*, Vol. 92 No. 2, pp. 493-511.

James, O., Jilke, S., Petersen, C. and Van de Walle, S. (2016), "Citizens' blame of politicians for public service failure: experimental evidence about blame reduction through delegation and contracting", *Public Administration Review*, Vol. 76 No. 1, pp. 83-93.

Janis, I.L. (1972), "Victims of groupthink: a psychological study of foreign-policy decisions and fiascoes", *The Journal of Politics*, Vol. 36 No. 1, pp. 218-277.

Jilke, S., Van Ryzin, G.G. and Van de Walle, S. (forthcoming), "Responses to decline in marketized public services: an experimental evaluation of choice overload", *Journal of Public Administration Research and Theory*.

Jilke, S., Van de Walle, S. and Kim, S. (2016), "Generating usable knowledge through an experimental approach to public administration", *Public Administration Review*, Vol. 76 No. 1, pp. 69-72.

Landis, J.R. and Koch, G.G. (1977), "The measurement of observer agreement for categorical data", *Biometrics*, Vol. 33 No. 1, pp. 159-174, available at: http://doi.org/10.2307/2529310.

*Landsbergen, D., Bozeman, B. and 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*, Vol. 2 No. 3, pp. 247-264.

Manjoo, F. (2013), "The happiness machine: how Google became such a great place to work", available at: www.slate.com/articles/technology/technology/2013/01/google_people_operations_the_secrets_of_the_world_s_most_scientific_human.single.html (accessed 12 January 2015).

Margetts, H.Z. (2011), "Experiments for public management research", *Public Management Review*, Vol. 13 No. 2, pp. 189-208.

Marvel, J.D. (2015), "Public opinion and public sector performance: are individuals' beliefs about performance evidence-based or the product of anti-public sector bias?", *International Public Management Journal*, Vol. 18 No. 2, pp. 209-227.

Morton, R.B. and Williams, K.C. (2010), *Experimental Political Science and the Study of Causality*, Cambridge University Press, New York, NY.

*Moynihan, D.P. (2013), "Does public service motivation lead to budget maximization? Evidence from an experiment", *International Public Management Journal*, Vol. 16 No. 2, pp. 179-196.

*Nielsen, P.A. and 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*, Vol. 25 No. 2, pp. 545-569. doi: 10.1093/jopart/mut051.

Niskanen, W.A. (1971), *Bureaucracy and Representative Government*, Aldine, Atherton and Chicago, IL.

*Nutt, P.C. (2005), "Comparing public and private sector decision-making practices", *Journal of Public Administration Research and Theory*, Vol. 16 No. 2, pp. 289-318.

Olsen, A.L. (2015), "Citizen (dis) satisfaction: an experimental equivalence framing study", *Public Administration Review*, Vol. 75 No. 3, pp. 469-478.

Open Science Collaboration (2015), "Estimating the reproducibility of psychological science", *Science*, Vol. 349 No. 6251, p. aac4716. doi: 10.1126/science.aac4716.

Perry, J.L. (2012), "How can we improve our science to generate more usable knowledge for public professionals?", *Public Administration Review*, Vol. 72 No. 4, pp. 479-482.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y. and Podsakoff, N.P. (2003), "Common method variance in behavioral research: a critical review of the literature and recommended remedies", *Journal of Applied Psychology*, Vol. 88 No. 5, pp. 879-903.

Raadschelders, J.C. and Lee, K.H. (2011), "Trends in the study of public administration: empirical and qualitative observations from public administration review, 2000–2009", *Public Administration Review*, Vol. 71 No. 1, pp. 19-33.

Riccucci, N.M., Van Ryzin, G.G. and Lavena, C.F. (2014), "Representative bureaucracy in policing: does it increase perceived legitimacy", *Journal of Public Administration Research and Theory*, Vol. 24 No. 3, pp. 537-551. doi: 10.1093/jopart/muu006.

*Scott, P.G. (1997), "Assessing determinants of bureaucratic discretion: an experiment in street-level decision making", *Journal of Public Administration Research and Theory*, Vol. 7 No. 1, pp. 35-58.

*Scott, P.G. and Pandey, S.K. (2000), "The influence of red tape on bureaucratic behavior: an experimental simulation", *Journal of Policy Analysis and Management*, Vol. 19 No. 4, pp. 615-633.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Company, Boston, MA.

*Sulitzeanu-Kenan, R. (2006), "If they get it right: an experimental test of the effects of the appointment and reports of UK public inquiries", *Public Administration*, Vol. 84 No. 3, pp. 623-653.

Tepe, M. (2015), "In public servants we trust?: a behavioural experiment on public service motivation and trust among students of public administration, business sciences and law", *Public Management Review*, pp. 1-31.

't Hart, P. (1990), *Groupthink in Government: A Study of Small Groups and Policy Failure*, Swets & Zeitlinger Publishers, Lisse, 318pp.

*Thurmaier, K. (1992), "Budgetary decisionmaking in central budget bureaus: an experiment", *Journal of Public Administration Research and Theory*, Vol. 2 No. 4, pp. 463-487.

Tranfield, D., Denyer, D. and Smart, P. (2003), "Towards a methodology for developing evidence-informed management knowledge by means of systematic review", *British Journal of Management*, Vol. 14 No. 3, pp. 207-222.

Tummers, L., Weske, U., Bouwman, R. and Grimmelikhuijsen, S. (2015), "The impact of red tape on citizen satisfaction: an experimental study", *International Public Management Journal*, pp. 1-22.

*Van De Walle, S. and Van Ryzin, G.G. (2011), "The order of questions in a survey on citizen satisfaction with public services: lessons from a split-ballot experiment", *Public Administration*, Vol. 89 No. 4, pp. 1436-1450.

*Van Ryzin, G.G. (2013), "An experimental test of the expectancy-disconfirmation theory of citizen satisfaction", *Journal of Policy Analysis and Management*, Vol. 32 No. 3, pp. 597-614.

*Walker, R.M., Brewer, G.a., Bozeman, B., Moon, M.J. and Wu, J. (2013), "An experimental assessment of public ownership and performance", *Public Management Review*, Vol. 15 No. 8, pp. 1208-1228. doi: 10.1080/14719037.2013.825480.

*Wittmer, D. (1992), "Ethical sensitivity and managerial decisionmaking: an experiment", *Journal of Public Administration Research and Theory*, Vol. 2 No. 4, pp. 443-462.

Further reading

*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*, Vol. 23 No. 3, pp. 631-661.

*Belle, 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*, Vol. 24 No. 1, pp. 109-136.

*Belle, N. and Cantarelli, P. (2015), "Monetary incentives, motivation, and job effort in the public sector: an experimental study with Italian government executives", *Review of Public Personnel Administration*, Vol. 35 No. 2, pp. 99-123.

*Bretschneider, S. and 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*, Vol. 2 No. 3, pp. 333-345.

*Burtraw, D., Goeree, J., Holt, C.A., Myers, E., Palmer, K. and Shobe, W. (2009), "Collusion in auctions for emission permits: an experimental analysis", *Journal of Policy Analysis and Management*, Vol. 28 No. 4, pp. 672-691. doi: 10.1002/pam.20460.

*Coursey, D.H. (1992), "Information credibility and choosing policy alternatives: an experimental test of cognitive-response theory", *Journal of Public Administration Research and Theory*, Vol. 2 No. 3, pp. 315-331.

*De Fine Licht, J. (2014), "Policy area as a potential moderator of transparency effects: an experiment", *Public Administration Review*, Vol. 74 No. 3, pp. 361-371.

*De Fine Licht, J., Naurin, D., Esaiasson, P. and Gilljam, M. (2014), "When does transparency generate legitimacy", *Experimenting on a Context-Bound Relationship. Governance*, Vol. 27 No. 1, pp. 111-134.

*Grimmelikhuijsen, S., Porumbescu, G., Hong, B. and Im, T. (2013), "The effect of transparency on trust in government: a cross-national comparative experiment", *Public Administration Review*, Vol. 73 No. 4, pp. 575-586. doi: 10.1111/puar.12047.

Grimmelikhuijsen, S.G. (2010), "Transparency of public decision-making: towards trust in local government?", *Policy & Internet*, Vol. 2 No. 1, pp. 5-35.

*Grimmelikhuijsen, S.G. (2012), "Linking transparency, knowledge and citizen trust in government: an experiment", *International Review of Administrative Sciences*, Vol. 78 No. 1, pp. 50-73.

*Grimmelikhuijsen, S.G. and 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*, Vol. 24 No. 1, pp. 137-157.

*Hock, S., Anderson, S. and Potoski, M. (2013), "Invitation phone calls increase attendance at civic meetings: evidence from a field experiment", *Public Administration Review*, Vol. 73 No. 2, pp. 221-228.

*Jakobsen, M. (2013), "Can government initiatives increase citizen coproduction? Results of a randomized field experiment", *Journal of Public Administration Research and Theory*, Vol. 23 No. 1, pp. 27-54.

*Jakobsen, M. and Andersen, S.C. (2013a), "Coproduction and equity in public service delivery", *Public Administration Review*, Vol. 73 No. 5, pp. 704-713.

*Jakobsen, M. and Andersen, S.C. (2013b), "Intensifying social exchange relationships in public organizations: evidence from a randomized field experiment", *Journal of Policy Analysis and Management*, Vol. 32 No. 1, pp. 60-82.

*James, O. (2011), "Managing citizens' expectations of public service performance: evidence from observation and experimentation in local government", *Public Administration*, Vol. 89 No. 4, pp. 1419-1435.

*Kaufmann, W. and Feeney, M.K. (2014), "Beyond the rules: the effect of outcome favourability on red tape perceptions", *Public Administration*, Vol. 92 No. 1, pp. 178-191.

*Knott, J.H. (2003), "Adaptive incrementalism and complexity: experiments with two-person cooperative signaling games", *Journal of Public Administration Research and Theory*, Vol. 13 No. 3, pp. 341-365.

*Landsbergen, D., Coursey, D.H., Loveless, S. and Shangraw, R.F. (1997), "Decision quality, confidence, and commitment with expert systems: an experimental study", *Journal of Public Administration Research and Theory*, Vol. 7 No. 1, pp. 131-158.

*Moynihan, D.P. (2005), "What do we talk about when we talk about performance? Dialogue theory and performance budgeting", *Journal of Public Administration Research and Theory*, Vol. 16 No. 2, pp. 151-168.

*Nam, Y., Kim, Y., Clancy, M., Zager, R. and Sherraden, M. (2013), "Do child development accounts promote account holding, saving, and asset accumulation for children's future? Evidence from a statewide randomized experiment", *Journal of Policy Analysis and Management*, Vol. 32 No. 1, pp. 6-33. doi: 10.1002/pam.21652.

Niskanen, W.A. (1994), "Bureaucracy and public economics", *John Locke Series in Classical Liberal Political Economy*, Edward Elgar Publishing, Inc.

*Van Ryzin, G.G. and Lavena, C.F. (2013), "The credibility of government performance reporting", *Public Performance & Management Review*, Vol. 37 No. 1, pp. 87-103.

*Vashdi, D.R. (2013), "Teams in public administration: a field study of team feedback and effectiveness in the israeli public healthcare system", *International Public Management Journal*, Vol. 16 No. 2, pp. 275-306.

*Wolf, P.J., Kisida, B., Gutmann, B., Puma, M., Eissa, N. and Rizzo, L. (2013), "School vouchers and student outcomes: experimental evidence from Washington, DC", *Journal of Policy Analysis and Management*, Vol. 32 No. 2, pp. 246-270. doi: 10.1002/pam.

Appendix

Theoretical principle	Operationalization	Values
Design and type	Of what specific type or subtype is the experiment?	(Basic design, crossover design, factorial design, longitudinal design, block design)
Contextual realism	Are subjects required to play a role during the experiment? (e.g. a public manager)	(Yes, No)
	What is the location of the experiment?	(Lab, classroom, field, online)
Participants	What types of subjects are used in the experiment? (e.g. practitioners or students)	(Students; practitioners; target population)
External validity	Are findings generalized?	(Yes, No)
		(Yes, No)
	Are findings generalized towards a specific (sub) population?	(Yes; some; not so much; not at all)
General design	Reported size of n	(R, ∞)
	Name of journal	(Name)

Table AI. Operationalization

About the authors

Robin Bouwman is a PhD Candidate Public Administration at the Institute for Management Research, Nijmegen, the Netherlands. His research interests are public sector negotiations, bargaining, behavioural public administration, lab experiments and experimental research methodology. Robin Bouwman is the corresponding author and can be contacted at: r.bouwman@fm.ru.nl

Stephan Grimmelikhuijsen is an Assistant Professor at the Utrecht University School of Governance, the Netherlands. He is interested in public sector transparency, trust, behavioural public administration and experimental methodology.