International Public Management Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/upmj20

Conducting Experiments in Public Management Research: A Practical Guide

Martin Baekgaard[a], Caroline Baethge[b], Jens Blom-Hansen[a], Claire A. Dunlop[c], Marc Esteve[d], Morten Jakobsen[a], Brian Kisida[e], John Marvel[f], Alice Moseley[c], Søren Serritzlew[a], Patrick Stewart[e], Mette Kjaergaard Thomsen[a] & Patrick J. Wolf[e]

[a] Aarhus University

[b] University of Passau

[c] University of Exeter

[d] University College London

[e] University of Arkansas

[f] George Mason University

Accepted author version posted online: 11 Mar 2015.

To cite this article: Martin Baekgaard, Caroline Baethge, Jens Blom-Hansen, Claire A. Dunlop, Marc Esteve, Morten Jakobsen, Brian Kisida, John Marvel, Alice Moseley, Søren Serritzlew, Patrick Stewart, Mette Kjaergaard Thomsen & Patrick J. Wolf (2015) Conducting Experiments in Public Management Research: A Practical Guide, International Public Management Journal, 18:2, 323-342, DOI: 10.1080/10967494.2015.1024905

To link to this article: http://dx.doi.org/10.1080/10967494.2015.1024905

International Public Management Journal

CONDUCTING EXPERIMENTS IN PUBLIC MANAGEMENT RESEARCH: A PRACTICAL GUIDE

MARTIN BAEKGAARD
AARHUS UNIVERSITY

CAROLINE BAETHGE
UNIVERSITY OF PASSAU

JENS BLOM-HANSEN
AARHUS UNIVERSITY

CLAIRE A. DUNLOP
UNIVERSITY OF EXETER

MARC ESTEVE
UNIVERSITY COLLEGE LONDON

MORTEN JAKOBSEN
AARHUS UNIVERSITY

BRIAN KISIDA
UNIVERSITY OF ARKANSAS

JOHN MARVEL
GEORGE MASON UNIVERSITY

International Public Management Journal, 18(2), pages 323–342
DOI: 10.1080/10967494.2015.1024905

ISSN: 1096-7494 print/1559-3169 online

ALICE MOSELEY
UNIVERSITY OF EXETER

SØREN SERRITZLEW
AARHUS UNIVERSITY

PATRICK STEWART
UNIVERSITY OF ARKANSAS

METTE KJAERGAARD THOMSEN
AARHUS UNIVERSITY

PATRICK J. WOLF
UNIVERSITY OF ARKANSAS

ABSTRACT: *This article provides advice on how to meet the practical challenges of experimental methods within public management research. We focus on lab, field, and survey experiments. For each of these types of experiments we outline the major challenges and limitations encountered when implementing experiments in practice and discuss tips, standards, and common mistakes to avoid. The article is multi-authored in order to benefit from the practical lessons drawn by a number of experimental researchers.*

INTRODUCTION

Observational data are routinely used in both quantitative and qualitative public management research. They can be used to study a broad range of research questions. However, it is often challenging to draw causal conclusions from such studies. This is due to omitted variables, reverse causality, and other endogeneity problems.[1] These problems may be difficult to avoid when collecting observational data. There are various techniques that can be employed *ex-post* to remedy them. But these solutions may not always be available, and they are often challenging in terms of complexity in both analysis and communication of results. In contrast, the core idea of experimental methods is to collect good data that do not need *ex-post* correction in order to be used for causal analysis. This is why experimental

methods are sometimes referred to as a design-based approach to causal research (e.g., Dunning 2012). As a result, the emphasis is on building a strong research design that collects good data. The quality of the data means that the ensuing analysis can often be done in a simple and transparent way. The analysis of the evidence may lie in a simple comparison of means between control and treatment groups. Experimental designs come in many different types with each having distinct advantages and disadvantages. These designs and their trade-offs are discussed in more detail by Blom-Hansen, Morton, and Serritzlew (2015). In this article, we focus on the practical challenges of applied experimental research. Our aim is to direct attention to these challenges, not to discuss them in depth. For a full treatment, we refer the reader to more specialized literature.[2]

Experiments may sound like a shortcut around the problems facing causal analysis of observational data. This is not necessarily true. Doing experimental research may sound easy, but this can be deceptive. Experimental researchers face a number of problems of their own. In the following sections, we discuss challenges often encountered when doing experiments in practice and provide advice on how to handle them. The article is multi-authored in order to benefit from the practical lessons drawn by a number of experimental researchers.

The article is structured into three sections where we discuss the practical aspects of, respectively, lab, field, and survey experiments. We focus on these three types of experiments because they are distinct, original and true experiments, and because, to date, they represent the most widely cited experiments in public management research. In each section, we provide a short description of the major challenges and limitations encountered when implementing experiments in practice, before discussing tips, standards, and common mistakes to avoid. We conclude the article by tying together the sections and discussing some unsettled questions about norms for experimental research within the field of public management.

LAB EXPERIMENTS IN PRACTICE

One of the first lab experiments was arguably a test of Galileo's uniform acceleration law. Here, the study was based upon two different balls—one made of lead and one of cork—rolling down an inclined plane. Through this experiment, Galileo revealed that no correlation exists between the different magnitudes of objects (i.e., their size and weight) and the speed at which they fall (Settle 1961). In other words, his experiment tested and supported the null hypothesis of no difference between conditions. Since then, lab experiments have been widely used in scientific disciplines across the world. When creatively conceived and executed with precision, lab experiments allow researchers to differentiate between the exogenous and the endogenous variables. They thus shed light on causal connections, allowing inferences to be made and conclusions to be drawn. However, lab experiments also suffer important limitations. In public management research, the main limitation is external validity. More precisely, the concern is whether the findings from lab settings can be applied to individual behavior in a range of organizational settings that are

encountered in the public management field. This particular aspect of external validity is sometimes referred to as ecological validity (Morton and Williams 2010, 264–265).

Despite their unique advantages, lab experiments in most social science disciplines are rare. Public management has largely neglected this methodological approach (for some notable exceptions, see James 2011; Knott, Miller, and Verkuilen 2003), despite scholars having acknowledged its importance (Bozeman and Scott 1992). Below, we provide practical suggestions concerning how lab experiments may be developed and implemented, as well as outlining common mistakes encountered while carrying out this type of study.

In order to cover the main areas that need to be planned when developing research through lab experiments, we use the UTOS framework (Units, Treatments, Observations, and Settings). Specifically, we rely on the definitive framework taken from Shadish, Cook, and Campbell (2002), which in turn builds upon the work of Cronbach. Cronbach states "each experiment consists of units that receive the experiences being contrasted, of the treatments themselves, of observations made on the units, and of the settings in which the study is conducted" (Shadish, Cook, and Campbell 2002, 19). We end the section on lab experiments by discussing common mistakes.

Units (Participants or Subjects)

The experimental subjects in lab experiments often consist of students. Student populations are easily recruited opportunity samples that are inexpensive or even "free" in the sense that participation in studies may be paid for in terms of course credit or extra credit (McDermott 2013). Further, some student populations may be theoretically interesting in themselves. For instance, Masters of Business Administration (MBA) and Masters of Public Administration (MPA) students often represent a mix of pre-service and in-service students that bring with them values, expectations, and experiences that are useful for understanding workplace interactions in private, non-profit, and public sectors (Dunlop and Radaelli 2014). However, in terms of generalizability, student populations have long been seen as potentially flawed (Sears 1986). Recent data comparing college student and adult populations find significant differences in personality traits (Stewart 2008). Thus, when considering students as subjects in lab experiments, researchers should elaborate on why this specific sample would be adequate for testing the hypothesis of their study (Morton and Williams 2010, 322–353).

Another easily accessed population is university employees. This population can reflect the general public to a great extent if the experimental sample is drawn from support staff (using professors as the sole study participants can obviously be quite problematic). Specifically, university staff members often reflect a broad range of personalities, values, educational attainment levels, and ages seen in the general public, and can be more easily recruited to take part in activities before, during, and after work hours.

Regardless of the type of subjects, researchers should carefully consider the question of anonymity. Subjects behave differently when their behavior can be observed

by the experimenter in person. To avoid this, neither the other participants nor the experimenter should be able to identify the subjects even if their actions/decisions (via computer) are observable. Anonymity is likewise important for subject recruitment and treatment assignment. Moreover, in most experimental designs researchers avoid possible sources of bias by randomly drawing subjects from a certain pool and/or randomly assigning them into different treatments. For example, if the experimenter systematically picks the subjects from a class and assigns them into treatments, they could adjust their behavior depending on what they think is socially desirable and expected by the experimenter. The same effect could be observed when participants know each other. This participant effect can be prevented by randomization and ensuring anonymity. The advantage of conducting computerized experiments is that the experimenter can ensure anonymity for the subjects who, in turn, cannot identify themselves via computer. This way, subject reaction to gender, age, appearance, social status, and other factors cannot bias behavior unless the experimenter provides this information to the subjects (presumably select information and for theoretical reasons).

Treatment

While Internet technology provides the opportunity to carry out studies from a distance, in-person studies allow for greater control of stimuli presentation. This can be exceptionally important when more emotional and viscerally impactful stimuli, such as visual images, smells, haptics (touching behavior), proxemics (personal proximity), and vocalics (voice tone), might play a role as either a treatment variable or as a potential confound to the treatment effect. Here, verisimilitude becomes important for public management research, as contextual elements play an important role in individual response.

Appreciating the pattern of an experimental effect is also important. An immediate first check is whether the treatment had the expected effect. This is followed by consideration of a treatment's latency (the interval between the stimulus and the response), how long the effect lasted, and the rate at which the treatment effect decays. At this point, researchers should consider several manipulation checks to ensure that no other variables can influence the effects of their independent variables on their dependent variables (Harris 2008).

The experimental instructions (either computerized or by pen and paper) should clearly state how the experiment will proceed, what the participants are expected to do, and, in the case of experimental games, whether there will be social interaction and with whom (partner, stranger or absolute stranger matching), and how the subject's payoff is affected by their own or others subjects' decisions. In order to ensure that the instructions are clearly understandable by the subjects, one or several pretests concerning the experiment with real participants should be conducted. Second, an ability or comprehension test should be included during the study before subjects start their experimental task. By including the test, the researcher can ensure that differential behavior cannot be attributed to either the subject's lack of understanding or misinterpretation of the instructions.

Finally, the researcher should endeavor to keep the amount and type of treatment modifications simple. If researchers include too many modifications they risk running out of degrees of freedom and being unable to draw any firm conclusions.

Observations

When putting together a study, an extensive literature review is necessary to avoid unnecessary replications of studies and to identify measures used in other closely related studies. Using previously employed measures gives greater confidence by providing a reference point and contributes to establishing cumulative knowledge.

Likewise, background measures should be collected using pre-existing questions/instruments that are theoretically relevant for the study's purposes. These measures should provide checks on the random assignment process, especially with smaller studies. Specifically, if theoretically interesting variables show statistically significant differences in the control and treatment groups, they may be entered in as covariates to control for potentially biasing characteristics.

The timing of the collection of background measures varies depending on the study itself. As a general rule, studies that exceed 15–20 minutes risk subject fatigue and burnout, which in turn might lead to "response sets" where subjects indicate choices in a systematic and automatic manner. If studies are relatively long, background variables may be collected either prior to, or after, a study takes place. However, connecting background data collected at an earlier stage with experimental data collected at a later time—or vice-versa—can be difficult if anonymity is to be maintained. A suggested strategy is to have subjects provide identifiers that are easily remembered by them, but are contextually confusing. For example, subjects can create a code based on different combinations of personal information, such as the first two letters of their mother's first name, the first two letters of their father's first name, the first two letters of their place of birth, the day of their birthday, or the last two letters of their mother's name. This would allow researchers to create a unique token for each of the individuals, while ensuring subject anonymity. On the other hand, if a study is relatively short, background information may be collected immediately prior to or after the experiment itself. In this case, care must be taken to not contaminate either the treatment or the background measures through the pre-test or post-test, respectively.

Setting

The final element of lab experiments is the setting. While this may vary from a normal classroom setting, in which a group of individuals respond to a stimulus, to a dedicated lab complete with computers devoted to the task of obtaining response information, the main benefit is the control it gives the researcher over external influences. Indeed, a level of creativity may be used with the study's setting so that greater authenticity may be had, which in turn enhances the generalizability of results. One option is to locate lab experiments in the subjects' natural environment rather than in university labs, so-called lab-in-the-field experiments

(Morton and Williams 2010, 296). There are several tools that can be used to conduct computerized experiments when not using a pen and paper setting. A commonly used tool for lab experiments is zTree by Fischbacher (2007). Other web-based tools that can either be used for lab or Internet experiments are BoXS (Seithe 2012), ComLab (Miller, Prasnikar, and Zupanic 2009), EconPort (Cox and Swarthout 2006), Multistage (Palfrey, Yuan, and Crabbe 2013), PEET (Saylor 2009), SoPhie (Hendriks 2012), and Willow (Weel and McCabe 2009). Some experimental programs also allow classroom experiments with access via mobile devices such as classEx (Giamattei 2014).

Common Mistakes

Due to the complexity of lab experiments, it is not uncommon that at the later stages of a research program—when researchers are analyzing their data or writing their results up—they realize their findings contradict previous theories that appeared to be well-established. This may be due to the very nature of the lab experiment, in which theory testing and replication across multiple sites advances the literature by pointing out flawed studies or the shortcomings of theoretical frameworks and hypotheses. However, it may just as likely be claimed that the experimental design did not consider certain key aspects of the theory, or was poorly implemented. While the former may be best avoided through a thorough literature review, the latter may be averted through careful implementation and rigorous oversight of the experimental study. To steer clear of this latter issue, we review common mistakes that researchers might encounter when carrying out lab experiments.

A first common mistake of many lab experiments is that they can be too complex, with too many variables and treatments affecting response to the dependent variable(s). Controlled lab experiments per se should not only include all necessary treatments but should especially make sure that subjects are not put off by the lab setting itself. The key to an impeccable experiment is simplicity. An epigram attributed to Albert Einstein states, "make everything as simple as possible, but not simpler." Simplicity in experimental studies can be achieved by including only a reasonable number of different treatments, by using understandable instructions, and by being implemented in a setting that is not so artificial as to be unsettling for the subject, or that interferes with their carrying out of the task at hand. In order to get unbiased results, it is also important to both control for subject and experimenter effects during the design and the experimental procedure. Double-blind procedures, in which both the subject and the researcher giving the treatment are unaware of whether it is a control condition or a treatment condition(s), are preferred to single-blind procedures, where only the subject is unaware if he or she is in the control or treatment group. This approach, in turn, minimizes the need for deception or to control for subject interpretation of the study (Christensen, Johnson, and Turner 2011).

An example of the effects of simplicity can be found in some lab experiments using behavioral economic theories, such as variations on the classic prisoner's dilemma game. As Andreoni (1995) argues, sometimes these frameworks are so complex that

most participants do not fully understand the game, and therefore the results of the experiment are misleading. A possible means to avoid such problems is to include questions at the end of each game or scenario asking participants to what degree they understood the setting, and the amount of mental effort they needed to engage with the experiment. By doing so, researchers can gauge whether or not participants fully comprehended the setting.

Another common mistake is forgetting to have a baseline treatment or a control group. This is necessary in order to be able to compare a treatment effect—i.e., by inter- or intra-group comparisons—and provides a useful means to test the null hypothesis that there is no treatment effect.

Furthermore, experiments should be incentivized correctly. If the design includes an economic game involving payoffs, the actual payoffs that subjects receive due to their specific decision within the experiment must be of a reasonable size; i.e., an average student wage per hour. If this is not taken into consideration, a subject's behavior may be distorted.

FIELD EXPERIMENTS IN PRACTICE

Field experiments apply random assignment techniques when implementing a policy, program, or administrative change. Because they happen in the real world, they have greater external validity than lab experiments, enhancing both their applicability and the power of policy arguments that may be made based upon the findings that result. At the same time, they also face many significant challenges (Gerber and Green 2012).

General Standards for Field Experiments

Field experiments require that access to an intervention be based on a random process such as a lottery. If study participants are selected for the treatment or control group based on their own decisions or the choices of other people, then experimental conditions are lost, and selection bias becomes a threat to the internal validity of the analysis. Similarly, if the randomization process itself is not sound, then compositional bias can confound comparisons between the outcomes of the treatment and control groups. Drawing from the model of medical drug trials, treatment group participants should receive a "dose" of the intervention large enough that we would expect it to make a difference in their outcomes. Ideally, the study participants who are randomly assigned to the control group experience "business-as-usual," consisting of whatever would have happened to them if the experimental intervention never existed. Finally, high-quality field experiments are informed by enough outcome data that allow researchers to detect substantively meaningful effects of the intervention when such effects actually exist. In other words, these treatments need to be adequately powered. The main challenges to successful field experiments generally fall within these four categories: the integrity of the random assignment, an adequate dosage of the treatment, the authenticity of the control group counterfactual, and data availability.

Ensuring the Integrity of the Randomization and Intervention

Randomization is challenging and can fail for a number of reasons, including a small sample size, multiple lotteries with inconsistent treatment assignment probabilities, or exceptions granted to particular study participants. Under such conditions, the treatment and control groups are likely to differ from each other regarding measurable or unmeasurable characteristics that could bias the resulting conclusions about the effectiveness of the intervention. An example might be the random assignment of elementary school students within a school to be taught by a teacher with special training (the treatment) or a teacher with conventional training (the control). There might be only a few dozen students in the experiment, different types of students may receive priority status based upon program goals and statutory guidance, or the principal at the school might bend to pressure from assertive parents who insist that their child be placed in the treatment classroom regardless of the lottery outcome. All of those circumstances would threaten to undermine the integrity of the randomization.

This challenge can be met in a number of ways. First, a positive working relationship should be established with staff at all levels of the organization implementing the intervention in order to encourage their cooperation with the research. In field experiments where researchers cooperate with a public organization or other large and/or complex organizations, setting up the experiment and implementing the intervention is, in many ways, demanding. It is not enough that the field experiment is approved at the political level. It must also be supported among the managers and street-level officials who will carry out the intervention. Second, a formal written agreement or a contract should be crafted with the implementing agency, which ensures that researchers will have the authority to design and implement the lottery and retain complete editorial control over reporting and publishing of study results. Third, sufficient time for participant recruitment should be allowed that will culminate in a single random assignment. Fourth, a test-randomization should be done prior to the actual randomization to confirm diagnostically that the randomization was successful. Diagnostics should also be run on the actual randomization. Fifth, if others perform the lottery, the researchers should convince the public authorities of the importance of the integrity of the random assignment, and should gather as much information as possible regarding how it was implemented. Sixth, if multiple lotteries are used, the participant observations should be weighted in the analysis by the inverse of their treatment assignment probability, thereby equating the two groups regarding any participant characteristics related to which lottery they were in. Finally, once participants are randomly assigned, they should stay in their assigned group regardless of whether or not they receive the treatment. Properly executed field experiments can have absolute control over this ''intention-to-treat,'' and often analyses based upon this initial assignment are the most policy relevant. Intention-to-treat analyses allow for the subjects to experience ''business as usual,'' with the exception of the treatment offer. Sometimes the treatment group will not comply with the offer of an intervention, and sometimes the control group will find a way to access the treatment. Yet, non-compliance on the part of the treatment or

control groups is not necessarily a problem since the aim of field experiments is often to measure the effect of an intervention as it would be if implemented in the actual administrative setting (including all the implementation challenges, non-compliance, and so on). Therefore, non-compliance is often of considerable practical interest since it informs researchers and practitioners about the implementation challenges a policy initiative based on the field experiment would face.

If withholding the treatment from the control group is undesirable or unethical, it may be possible to do a staged implementation of the treatment where the order in which different individuals or organizations receive the treatment is randomized. The individuals or organizations that receive the treatment at a later stage then function as a control group until they are treated (Kisida, Greene, and Bowen 2014).

Delivering an Adequate Dose of the Intervention

In public management, it is often difficult to know in advance how strong a "dose" of an experimental intervention would be expected to produce a clear effect on outcomes. If the intervention is providing customer or citizen information, the question is how much information is enough to obtain a significant effect? For example, if the intervention is a professional development workshop, what proportion of the treatment group has to attend and for how many days in order for their subsequent behaviors to change? Often, the absence of previous related studies puts the onus on researchers to make a best estimate.

This dosage estimation challenge can be met in two ways. First, the researcher should overcompensate when predicting how strong and sustained a dose of the intervention is necessary to generate an observable effect, as people tend to overestimate the efficacy of an exciting management intervention as well as the eagerness of treatment members to experience it. Second, the researcher should work closely with the street-level officials who will actually deliver the intervention to ensure it is being done with fidelity to the nature of the treatment and in a participant-friendly way that will encourage sustained exposure to the intervention.

Ensuring That the Control Group Experiences "Business-as-Usual"

The real world is messier and less predictable than a laboratory. People have free will and often use that freedom to make choices that threaten the efficacy of field experiments. For example, if a person agrees to participate in an evaluation of a professional development intervention, but are randomly assigned to the control group, they might sign up for an online professional development program on their own because the study recruitment process piqued their interest in enhancing their human capital. Another problem is the famed "Hawthorne" effect (Henderson, Whitehead, and Mayo 1937; but see Franke and Kaul 1978). If the study participants who randomly receive an intervention are more carefully studied by researchers than are members of the control group, the treatment members might generate different outcomes solely due to the influence of being watched.

This challenge can be met in three ways. First, the experiences of the treatment and control groups should be made as similar as possible in appearance, even while ensuring that the treatment participants actually receive the distinctive intervention. Second, the control group should never be denied an experience they would have had in the normal course of business simply because it is similar to the treatment being evaluated. Such experiences are part of the proper counterfactual. Third, the researcher should ensure that data collection protocols, including any direct observations of study participants, are similar between the treatment and control groups.

Collecting Enough Data

Many well-designed field experiments with sound randomizations are later undone due to insufficient data. Researchers might fall short of initial recruitment targets, fail to collect complete baseline data, suffer substantial study attrition that is greater in either the treatment or control group, or neglect to collect data on outcomes later deemed to be important to the study.

This challenge may be met in a number of ways. For this discussion, we define "baseline" to mean a specific point in time before study participants are randomized. First, since the baseline only happens once, the researcher should discuss prior to the project's launch what information should be collected before randomization, especially including baseline measures of the outcomes to be evaluated and critical participant demographic characteristics. Second, all baseline data (except perhaps permanent demographic information) should be collected prior to random assignment, as participant attitudes and behaviors might be altered immediately by assignment to the treatment or control group. Third, formal agreements should be established in advance that guarantee access to all the administrative or performance data that others will collect and which the researcher expects to require for the study; and researchers should continue to nurture those essential relationships throughout the project. Fourth, adequate project funds and time should be allocated to participant recruitment, and participation should be made as convenient and attractive as possible by writing data collection instruments in clear and highly accessible language and by staging data collection at convenient locations like participant workplaces, schools, or community centers. In addition, participants should be compensated, with cash if possible, in appropriate amounts for their time and trouble. Finally, data should be captured on all the key outcomes that are expected to be affected by the intervention so that the evaluation is complete.

Final Advice

We strongly recommend that researchers aspiring to conduct a field experiment have a backup plan. If randomization is not possible, due to low participation numbers or insufficient political buy-in from implementers, researchers should be prepared instead to implement the most rigorous quasi-experimental study possible. Although quasi-experiments tend to have less internal validity than experiments

(Blom-Hansen, Morton, and Serritzlew 2015), it is better to learn something about a public management intervention, with less certainty, than to learn nothing at all.

SURVEY EXPERIMENTS IN PRACTICE

Most of the practical challenges and problems that are associated with field and lab experiments also apply to survey experiments. However, some practical aspects are easier to deal with in survey experiments while others offer additional challenges. On the one hand, survey experiments are often easier to conduct in practice than field experiments because the researcher does not have to rely on the cooperation of political stakeholders. On the other hand, to a greater extent than field and lab experiments, survey experiments have to deal with issues of non-response. The main challenges we discuss here are maximizing the effectiveness of a treatment, questionnaire design, and sampling issues. We finish this section with a flavor of some recent and promising developments in survey experimental research.

Maximizing the Effectiveness of a Treatment

One important practical challenge when designing survey experiments is how to maximize the effectiveness of the experimental treatment. This should not be confused with whether the independent variable has the intended impact on the dependent variable (this is a question about whether the theoretical claim is supported or not), but has to do with whether the treatment is designed in a manner which makes the experiment a valid test of the theory (Mutz 2011, 86). This may not be the case if, for some reason, the treatment does not get through to the respondents or if, for instance, information presented in the treatment was known to respondents in the control group prior to the experiment.

Survey experimental researchers are therefore advised to conduct manipulation checks in order to test the effectiveness of the treatment. Such checks are typically based on one or more questions included in the survey in which the treatment also appears. Two examples may help illuminating these points. Van Ryzin (2013) randomly assigned respondents to receive low- or high-expectations statements from a government official and to view either low- or high-performance photographs in order to test whether expectations and performance perceptions have a causal impact on citizen satisfaction. Both experimental manipulations were checked by asking respondents simple questions about their expectations and performance perceptions after having received the manipulation. Baekgaard (2015), in a study of the causal impact of performance information on citizen service attitudes, compares the performance perceptions of respondents in the control and treatment group prior to presenting the treatment group with performance information in order to assess whether the two groups differ in their initial perceptions and knowledge of performance.

Question and Questionnaire Design

Analysts using survey experiments should not neglect standard issues of question and questionnaire design. Measurement validity and reliability are critical to the

design of any survey, including a survey in which one or more experiments are embedded. A number of factors bear on the validity and reliability of a survey question (or group of questions), including wording, the number of response options presented to subjects, and the labeling of those options. Excellent and accessible primers on question and questionnaire design are widely available (see, e.g., Krosnick and Presser 2010; Krosnick 1999), and so we do not belabor these issues here. Nevertheless, we encourage survey experimentalists to keep the following in mind: ''The heart of a survey is its questionnaire'' (Krosnick and Presser 2010, 263).

While non-experimental and experimental survey designs must attend to many of the same issues, it is important to note that one of these issues—question order—can be particularly consequential in the context of a survey experiment. More specifically, the proximity of a treatment variable(s) to an outcome variable(s) can be crucial for a survey experiment. Conventional wisdom is that treatment effects are amplified when the outcome measurement closely follows the treatment, and so placing multiple intervening questions between treatment and outcome can attenuate a treatment effect (Mutz 2011). Whether a treatment effect diminishes over time might be of substantive interest to a researcher, in which case building separation between treatment and outcome would be appropriate.

Multiple Experiments Embedded in a Single Survey

When considering whether (and how) to embed multiple experiments in a single survey, analysts must exercise common sense and, when available, look to theory and previous research for guidance. If it is plausible that one experiment's treatment might affect the way subjects respond to the second experiment's treatment, bundling both experiments in one survey should be done with care. For concreteness, consider a survey in which two performance information experiments are embedded. Both experiments are intended to examine how public sector performance information affects citizens' ratings of performance. One experiment focuses on how favorable information about the United States Postal Service affects citizens' general views of Postal Service performance. The second experiment focuses on how information about a local government's performance affects citizens' support for local government spending. In this scenario, the favorable information provided in the first experiment might predispose individuals to respond more favorably than they otherwise would to the information provided in the second experiment. Put differently, any positive attitudes generated about the Postal Service in the first experiment might spill over into the second experiment.

If analysts are intent on embedding two experiments in one survey (for cost reasons, perhaps), or in situations where they do not expect one experiment to confound another experiment, they should take steps to mitigate potential problems. Randomizing the order of the two experiments would be one step; another would be to place some distance between the two experiments. Additionally, analysts' decisions about embedding multiple experiments in one survey should be transparent. If an analyst presents the results from two bundled experiments in two separate outlets, the analyst should acknowledge doing so, and discuss how the bundling might have affected his or her results.

Sampling Issues

Samples can be gathered in a number of ways, either in person by researchers in field settings (James and Moseley 2014), using Internet panels (Moseley and Stoker 2015), or using other online recruitment tools such as MTurk, where participants are paid a nominal amount to complete a survey (Marvel 2015). They can also be administered directly to citizens by post or phone, using postal lists or telephone directories to identify participants (e.g., James 2011). The large sample size typical of survey experiments means they have high statistical power and therefore can detect small effect sizes. Survey experiments also provide an opportunity to collect additional information on covariates, something which can be difficult in field experiments (Mutz 2011). The breadth of information that can be potentially collected in survey experiments gives researchers greater scope for testing for heterogeneous effects on those with different demographic or attitudinal characteristics.

However, one has to be mindful that survey samples nearly always involve a self-selecting group of people who have agreed to take part. While survey companies and online mechanisms like MTurk are a reasonably good way of obtaining demographically representative samples (Buhrmester, Kwang, and Gosling 2011), respondents nevertheless may be atypical in other ways. They may, for example, have strong views about the topic of the survey, be opinionated or vocal people in general, or be motivated by payment. It is therefore important to take steps, whenever possible, to minimize respondent bias. For example, one can avoid stating the precise topic of the research at the initial recruitment stage to reduce participation by those with strong views.

Current Developments in Survey Experiments

Survey experiments are continually evolving. One important recent development in this area is the introduction of new methods for identifying causal mechanisms (Imai et al. 2011; Imai, Tingley, and Yamamoto 2013). These methods allow researchers to examine *why* changes in X cause changes in Y, in addition to merely testing *whether* changes in X cause changes in Y.

Since these methods have yet to be implemented in public management research, we consider an example from political science. Brader, Valentino, and Suhay (2008) ask *why* news about the costs of Latino immigrants increases white opposition to immigration more than news about the costs of European immigrants. Their theory is that news about Latino immigrants causes anxiety among whites, which in turn causes opposition to immigration. The effect of their treatment (news about Latino immigration costs), then, is mediated by anxiety. The problem for experimentalists is that randomly assigning individuals to different levels of a mediating variable like anxiety would seem to be impossible. However, as Imai et al. (2011; 2013) illustrate, survey experimentalists can test a mediation model like this by first randomly assigning subjects to a news condition, and then, within that condition, using priming techniques randomly to induce anxiety in some subjects but not others. In this approach, both the model's treatment and mediating variables are exogenously

varied (though because experimentalists cannot actually assign individuals to different anxiety levels, variation in anxiety will not be perfectly exogenous). Once randomization to treatment and mediating variables is done, an experimentalist would then use standard statistical tests to assess whether the effect of the treatment was direct, indirect, or some mixture of the two.

These types of methods could be useful for public management researchers who are interested in asking why citizens react to public sector performance information the way they do. Consider, for instance, the question of *why* citizens react negatively to unfavorable public sector performance information. Might citizens' negative reactions be emotional, and not purely rational? More specifically, might anger mediate the effect of unfavorable performance information on citizens' evaluations of public sector performance? This question becomes more interesting when the public sector is compared to the private sector: Does unfavorable public sector performance information make citizens angrier than unfavorable private sector performance information, and therefore cause citizens to evaluate public sector performance failures more harshly than private sector performance failures? Questions like these have practical public management implications. If citizens react angrily to public sector performance failures, public managers would not only need to come up with technical solutions to those failures. They would also need to come up with public relations strategies to allay any lingering citizen anger.

CONCLUSION

Experimental methods have clear strengths in their ability to address problems of omitted variables, reverse causality, and other endogeneity problems. This makes experimental methods valuable tools in public management research, where these problems are ubiquitous. Despite this, experimental methods are still little used in public management research, although that may be changing. A brief look in the abstracts of three leading public management journals shows that, in the first seven years of this millennium, only 33 articles mention experiments. In the next seven years, 47 did. In 2013 alone, 13 did.[3]

Experimental methods are still in their infancy in this discipline. Norms that are firmly established in related disciplines are not fully established in public management (cf. McDermott 2013). For instance, deception is close to taboo in experimental economics. It is seen by many as ethically problematic. Perhaps more importantly, it is also seen as disruptive to the credibility of experiments, and credibility is vital to many studies in experimental economics. In psychological research, on the other hand, deception is sometimes accepted, based on the view that important research questions simply cannot be answered otherwise. Another important norm relates to pre-registration of experimental protocols in official trial databases. Some experts argue that it is required to limit publication bias, others that this is an unnecessary obstacle. And, of course, many journals in other disciplines require that these different norms are respected. We have, in this article, refrained from taking sides on these issues. More experience with experiments in public management research is necessary before such questions can meaningfully be settled.

In summary, by providing practical advice on how lab experiments, field experiments, and survey experiments can be conducted in public management research, we have shown that experimental research can be quite simple when proper steps are taken. We hope that this can stimulate more experimental research in the future. We also hope that this practical advice will help researchers avoid some of the common pitfalls of experimentation in order to produce relevant and useful public management research.

NOTES

1. Sometimes endogeneity is understood only to mean reverse causality (e.g., King, Keohane, and Verba 1994, 185–196). But, as the introductory article to this special issue argues, endogeneity in a statistical model arises when an independent variable is correlated with the error term. This may arise not only because of reverse causality, but also because of measurement errors and omitted variables.
2. Much of this literature is cited in the rest of the article. For general introductions that also include practical guidance, we recommend Morton and Williams (2010) on lab experiments, Gerber and Green (2012) on field experiments, and Mutz (2011) on survey experiments.
3. Based on a SCOPUS search for "experiment*" in title, abstract, and keywords restricted to *International Public Management Journal, Journal of Policy Analysis and Management*, and *Journal of Public Administration Research and Theory* for the years 2000–2013.

REFERENCES

Andreoni, J. 1995. "Cooperation in Public-Goods Experiments: Kindness or Confusion." *The American Economic Review* 85(4): 891–904.

Baekgaard, M. 2015. "Performance Information and Citizen Service Attitudes: Do Cost Information and Service Use Affect the Relationship?" *International Public Management Journal* 18(2): 228–245.

Blom-Hansen, J., R. Morton, and S. Serritzlew. 2015. "Experiments in Public Management Research." *International Public Management Journal* 18(2): 151–170.

Bozeman, B. and P. Scott. 1992. "Laboratory Experiments in Public Policy and Management." *Journal of Public Administration Research and Theory* 2(3): 293–313.

Brader, T., N. A. Valentino, and E. Suhay. 2008. "What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat." *American Journal of Political Science* 52(4): 959–978.

Buhrmester, M., T. Kwang, and S. D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High Quality, Data?" *Perspectives on Psychological Science* 6(3): 3–5.

Christensen, L. B., B. R. Johnson, and L. A. Turner. 2011. *Research Methods, Design, and Analysis*, 11th ed. Boston: Pearson.

Cox, J. C. and T. J. Swarthout. 2006. "EconPort, Georgia State University." http://www.econport.org/content/experiments.html.

Dunlop, C. A. and C. M. Radaelli. 2014. "Teaching Regulatory Humility: Experimenting with Student Practitioners." Published electronically September 22, 2014. *Politics*. doi: 10.1111/1467-9256.12075.

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

Fischbacher, U. 2007. "z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Econonomics* 10(2): 171–178.

Franke, R. H. and J. D. Kaul. 1978. "The Hawthorne Experiments: First Statistical Interpretation." *American Sociological Review* 43(5): 623–643.

Gerber, A. S. and D. P. Green. 2012. *Field Experiments: Design, Analysis and Interpretation*. New York: W W. Norton and Company..

Giamattei, M. 2014. "classEx—Interactive Tool for Classroom-Experiments, Instructions and Software." http://www.wiwi.uni-passau.de/wirtschaftstheorie/classex-interaktive-hoersaalexperimente/

Harris, P. 2008. *Designing and Reporting Experiments in Psychology*, 3rd ed. Berkshire, UK: Open University Press.

Henderson, L. J., T. N. Whitehead, and E. Mayo. 1937. "The Effects of Social Environment." Pp. 143–159 in L. Gulick and L. Urwick, eds., *Papers on the Science of Administration*. New York: Institute of Public Administration.

Hendriks, A. 2012. "SoPHIE: Software Platform for Human Interaction Experiments." Working Paper. http://www.sophie.uni-osnabrueck.de/

Imai, K., L. Keele, D. Tingley, and T. Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4): 765–789.

Imai, K., D. Tingley, and T. Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society*, Series A 176(1): 5–51.

James, O. 2011. "Performance Measures and Democracy: Information Effects on Citizens in Field and Laboratory Experiments." *Journal of Public Administration Research and Theory* 21(3): 399–418.

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

King, G., R. O. Keohane, and S. Verba. 1994. *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Kisida, B., P. Greene, and D. H. Bowen. 2014. "Creating Cultural Consumers: The Dynamics of Cultural Capital Acquisition." *Sociology of Education* 87(4): 281–295.

Knott, J. H., G. J. Miller, and J. Verkuilen. 2003. "Adaptive Incrementalism and Complexity: Experiments with Two Person Cooperative Signaling Games." *Journal of Public Administration Research and Theory* 13(3): 341–365.

Krosnick, J. A. 1999. "Survey Research." *Annual Review of Psychology* 50(1): 537–567.

Krosnick, J. A. and S. Presser. 2010. "Question and Questionnaire Design." Pp. 263–314 in P. V. Marsden and J. D. Wright, eds., *Handbook of Survey Research*, 2nd ed. Bingley, UK: Emerald.

Marvel, J. 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* 18(2): 209–227.

McDermott, R. 2013. "The Ten Commandments of Experiments." *PS: Political Science* 46(3): 605–610.

Miller, R. A., V. Prasnikar, and D. Zupanic. 2009. "ComLabGames, Technical Documentation." http://www.comlabgames.com.

Morton, R. B. and K. C. Williams. 2010. *Experimental Political Science and the Study of Causality: From Nature to the Lab*. Cambridge: Cambridge University Press.

Moseley, A. and G. Stoker. 2015. "Putting Public Policy Defaults to the Test: The Case of Organ Donor Registration." *International Public Management Journal* 18(2): 246–264.

Mutz, D. C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.

Palfrey, T. R., W. Yuan, and C. Crabbe. 2013. "Multistage: A Framework for Rapid Development of Experimental Games." SSEL, Caltech and CASSEL UCLA. http://multistage.ssel.caltech.edu:8000/multistage/.

Saylor, B. 2009. "PEET: UAA Python Experimental Economics Toolkit." University of Alaska Anchorage. https://econlab.uaa.alaska.edu/Software.html.

Sears, D. O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3): 515–530.

Seithe, M. 2012. "Introducing the Bonn Experiment System (BoXS)." Bonn Econ Discussion Papers No. 01/2012. http://boxs.uni-bonn.de/.

Settle, T. B. 1961. "An Experiment in the History of Science." *Science* 133: 19–23.

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

Stewart, L. E. 2008. "College Students as Research Subjects: Are Study Results Generalizable?" Poster for PsyPAG Annual Conference, University of Manchester, UK, July 30–August 1.

Van Ryzin, G. G. 2013. "An Experimental Test of the Expectancy-Disconfirmation Theory of Citizen Satisfaction." *Journal of Policy Analysis and Management* 32(3): 597–614.

Weel, J. and K. McCabe. 2009. "Willow: Experiments in Python." George Mason University. http://econwillow.sourceforge.net/.

ABOUT THE AUTHORS

Martin Baekgaard (MartinB@ps.au.dk) is an Associate Professor at the Department of Political Science, Aarhus University. His current research interests are performance information use and political-administrative relations. His research has been published in journals like *Journal of Public Administration Research* and *Theory, Governance, and Public Administration*.

Caroline Baethge (Caroline.Baethge@Uni-Passau.De) is a Research Assistant, Chair for Management, People and Information at the University of Passau, Germany.

Jens Blom-Hansen (jbh@ps.au.dk) is a Professor in the Department of Political Science, Aarhus University, Denmark. His research interests include public administration and EU politics.

Claire A. Dunlop (c.a.dunlop@exeter.ac.uk) is Associate Professor in Political Science at the Department of Politics, University of Exeter, UK. Her research interests cover regulation, policy analysis and the science-policy interface. Most recently, she has published research articles in *Political Studies, Regulation and Governance, Journal of European Public Policy, Policy Studies* and *Policy Sciences.* Claire co-edits *Public Policy and Administration*.

Marc Esteve (marc.esteve@ucl.ac.uk) is an Assistant Professor in the School of Public Policy at the University College London. His current research focuses on the role of core personality variables on the outcomes of collaborations by using experimental designs.

Morten Jakobsen (mortenj@ps.au.dk) is an Assistant Professor at the Department of Political Science at Aarhus University. His research interests include coproduction of public services, the relationship between public administration and citizens, public employees, information and communication in bureaucracy, and various forms of political participation.

Brian Kisida (bkisida@uark.edu) is a Senior Research Associate in the Department of Education Reform at the University of Arkansas. His academic work emphasizes experimental methodologies and includes studies of school governance and school choice, school integration, informal learning, cultural institutions, and art and music education.

John Marvel (jmarvel@gmu.edu) is an Assistant Professor in the School of Policy, Government, and International Affairs at George Mason University. His research focuses on public management issues, including public sector employee turnover, public sector work motivation, and manager-employee relationships in public sector organizations. He also does work on public opinion, focusing on how citizens react to public- and private-sector performance failures and successes differently.

Alice Moseley (A.Moseley@ex.ac.uk) is a Research Fellow in the Department of Politics at the University of Exeter, where she received her PhD in politics. Her current research interests include behavior change and coproduction in public policy, institutions for solving collective action problems, and the organization and governance of public services.

Søren Serritzlew (soren@ps.au.dk) is a Professor of Political Science at Aarhus University, Denmark. His research interests include effects of public sector reform, use of economic incentives in the public sector, and democracy.

Patrick A. Stewart (pastewar@uark.edu) (PhD in Political Science, Northern Illinois University) is an Associate Professor in the Department of Political Science at the University of Arkansas, Fayetteville. He is interested in nonverbal communication in the context of workplace and political interactions. His books include *Debatable Humor: Laughing Matters on the 2008 Presidential Primary Campaign* and *The Invisible Hands of Political Parties in Presidential Elections*.

Mette Kjærgaard Thomsen (mkt@ps.au.dk) is a doctoral student in the Department of Political Science and Government at Aarhus University. She is currently working on her PhD thesis on the effect of different types of government initiatives on citizen input to coproduction.

Patrick J. Wolf (pwolf@uark.edu) is Distinguished Professor and 21st Century Endowed Chair in School Choice at the University of Arkansas in Fayetteville. His specialty is conducting rigorous empirical evaluations of policy reforms. He has authored, co-authored, or co-edited four books and over 100 journal articles, book chapters, and policy reports on school choice, civic values, public management, special education, and campaign finance. He received his PhD in Political Science from Harvard University in 1995.