Version: 19 February 2014

Doing Experiments in Public Management Research: A Practical Guide

Authors:
Caroline Baethge,Faculty of Business Administration and Economics, University of Passau, Germany (email: caroline.baethge@uni-passau.de)
Jens Blom-Hansen, Department of Political Science, Aarhus University, Denmark (email: jbh@ps.au.dk)
Martin Bækgaard, Department of Political Science, Aarhus University, Denmark (email: MartinB@ps.au.dk)
Claire A. Dunlop, Department of Politics, University of Exeter, UK (email: C.A.Dunlop@exeter.ac.uk)
Marc Esteve, School of Public Policy, University College London, UK (email: marc.esteve.laporta@gmail.com)
Morten Jakobsen, Department of Political Science, Aarhus University, Denmark (email: MortenJ@ps.au.dk)
Brian Kisida, Department of Education Reform, University of Arkansas, USA (email: briankisida@gmail.com)
John D. Marvel, Department of Public & International Affairs, George Mason University, USA (email: jdmarvel@gmail.com)
Alice Moseley, Department of Politics, University of Exeter, UK (email: A.Moseley@exeter.ac.uk)
Søren Serritzlew, Department of Political Science, Aarhus University, Denmark (email: soren@ps.au.dk)
Patrick A. Stewart, Department of Political Science, University of Arkansas, USA (email: pastewar@uark.edu)
Mette Kjærgaard Thomsen, Department of Political Science, Aarhus University, Denmark (email: mkt@ps.au.dk)
Patrick J. Wolf, Department of Education Reform, University of Arkansas, USA (email: pwolf@uark.edu)

Doing Experiments in Public Management Research: A Practical Guide

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 studies of observational data. This is due to selection bias, omitted variables and endogeneity, which are problems that may be difficult to avoid when collecting observational data. There are various techniques that can be employed *ex-post* to remedy these problems. But these solutions may not always be available, and they are often challenging in terms of complexity. In contrast, the core idea of experimental methods is 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. The emphasis is on building a strong research design. The quality of the data means that the ensuing analysis of the collected data can often be done in a simple and transparent way. The evidence may lie in a simple comparison of means between control and experiment groups. Experiments come in different types which each have distinct advantages and disadvantages. They are discussed in more detail by Blom-Hansen, Morton and Serritzlew (2015). In this paper we focus on the challenges of practical experimental research.

Experiments may sound like a short cut around the problems facing causal analysis of observational data. However, 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 we discuss challenges often encountered when doing experiments in practice and provide advice on how to handle them. The paper is multi-authored in order to benefit from the lessons drawn by a number of experienced experimental researchers.

The paper is structured in 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 pure experiments and because, so far, they represent the most widely used 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 end the paper by a

conclusion where we tie things together and point to future avenues of experimental research within public management research.

Lab experiments in practice

Galileo Galilei's uniform acceleration law was arguably the first theory to be tested using a lab experimental design. 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 specific magnitudes of objects (i.e., their size and weight) and the different speeds if they fall over long spaces (Settle 1961). Since then, lab experiments have been widely used in several scientific disciplines across the world. When creatively conceived and executed with precision, 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. The main one is ecological validity, or in other words, if 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.

Despite the unique advantages, lab experiments in most social science disciplines are highly uncommon. Public management has seemingly deliberately neglected this methodological approach (see, for some notable exceptions, James 2011; Knott, Miller & Verkuilen 2003), despite scholars having acknowledged its need (Bozeman & Scott 1992). Below, we provide practical suggestions concerning how lab experiments may be developed and implemented, as well as reveal 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. Here, 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 & Campbell 2002: 19). We end the discussion of lab experiments by discussing common mistakes.

Units (participants or subjects)

Student samples are useful for testing theoretical connections, although in terms of generalizability, this "college sophomore" population has long been seen as potentially flawed (Sears 1986), with recent data comparing college student and adult populations and finding significant differences in personality traits (Stewart 2008). However, 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. 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. Thus, when considering undergraduate or graduate students as the population of an experiment, researchers should elaborate on why this specific sample would be adequate for testing the hypothesis of their study.

Another easily accessed population is that of 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 study participants can obviously be quite problematic). Specifically, university staff members often reflect a broad range of personalities, values, educational attainment, ages, and sex that is seen in the general public of a region, and can be more easily recruited to take part in activities before, during, and after work hours.

Regardless of the type of participants, researchers must ensure anonymity. Subjects behave differently when their behavior can be observed by the experimenter in person (single-blindness). If this is not part of the research question the design should involve a double-blind procedure, that is, neither the other participants nor the experimenter can identify the subject even if its 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 biases by randomly drawing subjects from a certain pool 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 etc. cannot bias behavior unless the experimenter provides this information to the interacting 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 emotionally relevant and viscerally impactful stimuli such as visual images (Grabe & Bucy 2009; Stewart, Salter, & Mehu 2009), 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. 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) and how long the effect lasted, and the rate at which the treatment effect delays. 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. Secondly, an ability or comprehension test should preferably 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 treatments modifications simple. If researchers include too many modifications they run the risk of not only wasting money and time, but also mis-answering the actual research question.

Observations

When putting together a study, an extensive literature review is necessary to not only make sure that unnecessary replications of studies are not carried out, but also to find measures from other closely related studies. Using these measures, with similar or same ranges of response, provides confidence in response by providing a reference point. It also helps to understand populations when the statistical response means, medians, modes, and standard deviations are different by potentially identifying reasons within the study's sample. It also helps to identify problems with the measures themselves that might manifest in ceiling or floor effects.

Likewise, background measures should be collected using pre-existing questions/instruments that are theoretically interesting 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(s) groups, they may be entered in as covariates to control for potentially biasing effects.

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 that is not thoughtful or insightful. If studies are relatively long, background variables may be collected either prior to or after a study takes place. However, connecting the files together 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 mother's first name, first two letters of father's first name, first two letters of place of birth, the day of their birthday, or the last two letters of 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 setting a study takes place in so that greater authenticity may be had, which in turn enhances the generalizability of results.

Common mistakes

Due to the complexity of lab experiments, it is not uncommon that at the later stages of a research programme – when researchers are analyzing their data or writing their results up – that 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. However, the key to an impeccable experiment is simplicity. As stated by Albert Einstein, "make everything as simple as possible, but not simpler." Simplicity in experimental studies should include a reasonable number of different treatments/ experimental groups/ modified variables, understandable instructions, and a setting that is not so artificial as to be unsettling for the subject, and interfere with their carrying out of the task at hand. In order to get unbiased results it is furthermore important to both control for subject and experimenter effects during the design and the experimental procedure, that is, 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, whereas 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 & Turner 2011). An example of the effects of simplicity can be found in some lab experiments using behavioral economic settings, 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 that they need to engage with the experiment. By doing so, researchers can ensure that participants fully comprehended the setting.

Other common mistakes include 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, the 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 subject's behavior can be distorted.

Field experiments in practice

Field experiments apply random assignment techniques when actually implementing a policy, program, or administrative change. Because they happen in the real world, they have greater external validity than lab experiments but they also face many significant challenges (Gerber & 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 or 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 the 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 a sufficient amount of outcome data that they can detect substantively meaningful effects of the intervention when such effects actually exist. In other words, they need to be adequately powered. The main challenges to successful field experiments generally fall within these 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

The challenge is that a randomization 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 students within a school to be taught by a teacher with a special certification (the treatment) or a teacher with a conventional certification (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 to encourage their cooperation with the research. It is not enough that the field experiment is approved at the political level, it must also be supported among managers and street-level bureaucrats who carry out the intervention. Second, a formal written agreement or a contract should be crafted with the implementing agency that ensures that researchers will have the authority to design and implement the lottery and retain complete editorial control over the reporting and publishing of the 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 and diagnostics should be run on the two randomized groups to confirm that the

randomization was successful. Such diagnostics also should be run on the actual randomization. Fifth, if others perform the lottery, the researchers should convince 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. 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. Non-compliance on the part of the treatment or control groups is acceptable so long as it represents real-world conditions.

Delivering an adequate dose of the intervention

The challenge is that 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 customer information, how much intervention is enough? 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 an effect to be clear in their subsequent behaviors? Often we can only guess.

This challenge can be met in two ways. First, the researcher should be liberal in guessing how strong and sustained of 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 bureaucrats who will actually deliver the intervention to ensure that 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 challenge is that 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 purity of field experiments. For example, if a person agrees to participate in an evaluation of a professional development

intervention, but they are randomly assigned to the control group, they might sign up for an on-line 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 & Mayo 1937). If the study participants who randomly receive an intervention are more carefully studied by researchers than are the members of the control group, the treatment members might generate different outcomes solely due to the social pressure 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

The challenge is that 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. First, since baseline only happens once, the researcher should brain-storm prior to project launch regarding what information should be collected at baseline, 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. Fourth, sufficient time and effort should be invested into maintaining constructive relationships with program implementers to ensure the continuous flow of information. Fifth, 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. Sixth, participants should be compensated, with cash if possible, in appropriate amounts for their time and trouble. Seventh, the location of study participants should be tracked carefully and frequently, by asking for contact info at baseline for themselves, other close family members, and a neighbor, and then they need to be regularly re-contacted to determine if they have changed addresses, telephone numbers, or email addresses. 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

Finally, we 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 to instead implement the most rigorous quasi-experimental study possible. Although quasi-experiments tend to have less internal validity than experiments, it is better to learn something about a public management intervention, with less confidence, 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. Moreover, survey experiments often offer excellent opportunities to test the experimental treatment in advance of the experiment due to a large and heterogeneous population of respondents.

On the other hand, to a greater extent than field and lab experiments, survey experiments have to deal with issues of non-response. This has to do with survey experiments being conducted under artificial conditions and often with the use of a highly heterogeneous group of respondents (citizens recruited through internet panels, for instance) who have little or no interest in the research being conducted.

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 right or wrong), but has to do with whether the treatment is effective in producing the intended variation in the independent variable (Mutz 2011: 86). Such variation is an important precondition for testing the theory. The manipulation should produce variation in the intended direction in the independent variable for the experiment to be a valid test of the theory.

Survey experimental researchers are therefore advised to do 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. Manipulation checks may either be used to test whether the treatment got through to respondents in the treatment condition or whether information presented in the treatment was known to respondents in the control condition prior to the experiment.

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 & 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 & 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.

Sampling issues

Samples can be gathered in a number of ways, either in person by researchers in field settings (James & Moseley forthcoming), using internet panels (Moseley & Stoker 2015), or using other online recruitment tools such as MTurk, where participants are paid a nominal amount to complete a survey (Marvel & Pitts 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 when these are present. 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 – see for example Buhrmester, Kwang and Gosling (2011) on MTurk samples – 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 minimise 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 (see Moseley & Stoker 2015).

External validity

Survey experiments' external validity is highest when they are conducted with representative and heterogeneous samples (rather than merely with student samples, for example) and in real world settings (such as within a workplace, in a home, or during use of a public service). Real world settings make participants' behaviour more natural than in a lab, and the 'noise' of the more naturalistic setting means that the treatments are being received in a context that is more like one where a policy or public management intervention might be conducted. One should exercise caution in claims about the external validity of a single survey experiment. Ideally, survey experiments conducted in the field should be replicated in several field settings before claims about generalizability can start to be made.

The external validity of survey experiments can be further enhanced when the treatments use real information – e.g., real performance information, as with Bækgaard (2015); see also James (2011). Lastly, external validity can be enhanced by using different outcome measures within one survey to operationalise the same concept. A finding that is replicated across different versions of the outcome measure is more likely to be externally valid (Mutz 2011).

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, Keele, Tingley & Yamamoto 2011; Imai, Tingley & 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, 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

immigration costs), then, is mediated by anxiety. Imai et al. (2013; 2011) illustrate how 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 to randomly induce anxiety in some subjects but not others. 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.

Using old-fashioned random assignment in concert with new technologies, researchers are pushing the boundaries of the survey experiment. Grimmer, Messing & Westwood (2012), for instance, use Facebook as the platform for a survey experiment about the effects of legislators' credit-claiming (for spending) on citizens' credit allocations. Given that many individuals communicate about politics via Facebook, Grimmer et al.'s (2012) use of this popular platform increases the external validity of their study. More generally, their study suggests that when it comes to designing survey experiments, researchers are limited only by their imaginations.

Conclusion

Experimental methods have clear strengths in their ability to address problems of selection bias, omitted variables and endogeneity. 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 they seem to be on the rise. A brief look in the abstracts of three leading journals shows only that, in the first seven years of this millennium, only 33 articles mention experiments. In the next seven years, 47 did. In 2013, 13.[1]

We hope, by providing some practical advice on how lab experiments, field experiments and survey experiments can be conducted in public management research, to have shown that experimental research can be quite simple and easy, and that this can stimulate more experimental research in the future. We also hope that the practical advice will help researchers avoid some of the common pitfalls.

Experimental methods are thus still in its infancy in this discipline. This means that norms, that are firmly established in related disciplines, are not yet here. 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

[1] Based on a SCOPUS search for "experiment*" in title, abstract and keywords restricted to this journal, *Journal of Policy Analysis and Management*, and *Journal of Public Administration Research and Theory* for the years 2000-2013.

many studies in experimental economics. In psychological research, on the other hand, deception is sometimes accepted. Important research questions simply cannot be answered without. Another important norm relates to pre-registration of experimental protocols. Some 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 in these discussions. More experience (and experiments) with experiment in public management research is in our opinion necessary, before such questions can be meaningfully settled.

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, Cost Information, and Citizen Service Attitudes. *International Public Management Journal*.

Blom-Hansen, J., Morton, R. & Serritzlew, S. (2015). Experiments in Public Management Research. *International Public Management Journal*.

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

Brader, T., Valentino, N. A., & Suhay, E. (2008). What triggers public opposition to immigration? Anxiety, group cues, and immigration threat. *American Journal of Political Science*, *52*(4): 959–978.

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

Christensen, L. B., Johnson, B. R. & Turner, L. A. (2011). *Research Methods, Design, and Analysis*. Eleventh Edition. Boston: Pearson.

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

Grabe, M. E., & Bucy, E. P. (2009). *Image Bite Politics: News and the Visual Framing of Elections*. New York: Oxford University Press.

Grimmer, J., Messing, S., & Westwood, S. J. (2012). How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation. *American Political Science Review*, 106(4): 703–719.

Harris, Peter. (2008). *Designing and Reporting Experiments in Psychology*. Third Edition. Berkshire, England: Open University Press, McGraw-Hill.

Henderson, L. J., Whitehead, T. N., & Mayo, E. (1937). The Effects of social environment. In Gulick, L., & Urwick, L. (ed.), *Papers on the science of administration*. New York: Institute of Public Administration.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (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., Tingley, D., & Yamamoto, T. (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 & Theory*, 21(3): 399-418.

James, O. & Moseley, A. (forthcoming). Does Performance Information about Public Services Affect Citizens' Perceptions, Satisfaction and Voice Behaviour? Field Experiments with Absolute and Relative Performance Information. *Public Administration*.

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

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

Krosnick, J. A., & Presser, S. (2010). *Question and questionnaire design. Handbook of Survey Research*. Second edition. Bingley, UK: Emerald, 263–314.

Marvel, J. & Pitts, 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*.

Moseley, A. & Stoker, G. (2015). Putting public policy defaults to the test: A survey experiment in organ donor registration. *International Public Management Journal*.

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

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.

Settle, T. B. (1961). An Experiment in the History of Science. *Science*, 133: 19-23.

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

Stewart, P. A., Salter, F. K., & Mehu, M. (2009). Taking leaders at face value: Ethology and the analysis of televised leader displays. *Politics and the Life Sciences*, 28(1): 48-74.

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