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

Experiments in Public Management Research

Jens Blom-Hansen[a], Rebecca Morton[b] & Søren Serritzlew[a]
[a] AARHUS UNIVERSITY
[b] NEW YORK UNIVERSITY

Accepted author version posted online: 11 Mar 2015.

EXPERIMENTS IN PUBLIC MANAGEMENT RESEARCH

Jens Blom-Hansen[1], Rebecca Morton[2], Søren Serritzlew[1]

Jens Blom-Hansen (jbh@ps.au.dk) is in the Department of Political Science, Aarhus University, Denmark.

Rebecca Morton (rebecca.morton@nyu.edu) is a Professor of Politics at New York University, USA. Her research interests include American elections, empirical analysis of formal models of politics, and experimental methods.

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.

[1]AARHUS UNIVERSITY, [2]NEW YORK UNIVERSITY

Abstract

This paper argues that experimental methods are underused in public management research. This is lamentable, since this research field faces especially severe endogeneity problems. We introduce five different experimental designs together with a discussion of their strengths and weaknesses in public management research: Lab, survey, field, natural, and quasi-experiments. We also discuss whether experiments are low on external validity. This objection is often raised, but we think it is false.

INTRODUCTION

One of the thorniest problems in public management research is to identify the causal effect of key independent variables. For example, organizational reforms are typically made in response to existing problems. This reality renders it difficult for empirical research to disentangle the effects of organizational changes from the reasons for making such changes.

From a practical research design perspective, this question is crucial. Organizational reform and changes in performance may be correlated, but reform only brings about improved performance if the relationship is causal. Methodologically speaking, the root of this problem is often policy endogeneity (Besley and Case 2000), which occurs when policy changes are made in response to the factors that the policies are designed to affect. For instance, a reform could be implemented in a certain area because of poor performance, and in the hope that the reform will remedy the problem. However, if the areas where reforms are not implemented are areas where performance is acceptable, then it is impossible to know whether differences in performance between the two areas are due to the reform – or to prior differences in performance. Generally, something is endogenous to a system when it is determined within the system. To take a concrete example, reformers may introduce private competition in some schools because they want to pressure these schools to increase performance. So the perceived need for reform (low performance) is the cause of the choice of reform (private competition). Any correlation between reform and performance will then be biased.

This problem is hardly surprising. Civil servants are, after all, hired to solve specific problems—or, in a manner of speaking, to create problems of policy endogeneity for researchers. Consequently, endogeneity problems are particularly relevant for this line of research. They are likely to appear whenever we study the effects of manipulatable aspects of the public sector. Studies relying on traditional observational data are therefore likely to suffer from biased estimates of the impact of organizational reforms. While some studies discuss endogeneity problems, it is no exaggeration that public management

research commonly neglects these issues. Meta-analyses and literature surveys of the field regularly complain of a lack of methodological rigor.

The classic solution to endogeneity problems is experimental methods. As explained below, experimental methods can make the reform exogenous by randomly assigning reforms and also provide the counterfactual observations of what happens when reforms do not take place.

Experimental methods have become mainstream in economics and increasingly popular in political science. However, experimental methods are still underused in public management research despite their great potential.

The aim of this paper is, thus, to argue that experiments are underused in public management research. The problem of endogeneity and why it is particularly relevant for public management research is discussed, after which experiments are argued to offer design-based solutions to the problem. Finally, five different experimental designs are presented together with discussion of their strengths and weaknesses. In the conclusion, we discuss whether experiments are low on external validity. This is an objection that is often heard, but one which we think is flawed.

Endogeneity—What Is It?

Formally, a variable is said to be endogenous when is it correlated with the error term of the model (Wooldridge 2013, 83). Standard statistical methods are biased when

endogeneity is present. In general, the problem of endogeneity has several sources. Two typical problems in public management research are self-selection and simultaneity. Selection problems are likely to be serious whenever a choice that is studied is a consequence of the outcome of the choice. Consider, for example, the problem confronting a researcher who would like to study the effects of military service on the wages of workers after having completed their service. When military service is voluntary, some may choose the military because of the possible education and training benefits for later use in other occupations. If those choosing military service are more likely to benefit from such training (e.g., because they are more motivated or have greater aptitude), then discovering that those who have served in the military earn higher wages when they leave may not fully reflect the effect of military service. In other words, these individuals might have secured the training anyway and earned higher wages even if they had not served in the military. Alternatively, because they have higher aptitude, they may earn higher wages because they have better innate abilities. Discovering that military service leads to higher wages may have nothing to do with the military service, but instead with the types of individuals who voluntarily choose to join the military and later leave for nonmilitary jobs.

Because military service in this example is voluntary, individuals select whether they are "treated" or not (serve in the military), and there are factors (motivation, aptitude) determining whether they are treated, which also affect the outcome variable under investigation (wages outside the military). It is therefore a self-selection problem. Ideally, we would want the probability of serving in the military to be independent of these

factors in order to measure the effects of military service on wages. Instead, since selection into military service in order to seek training and education, in this example, is a function of the outcome variable we are interested in (wages outside the military), we cannot fully distinguish the effects of military service on these wages.[1]

Simultaneity is sometimes referred to as "two-way causation" or reverse causation. For instance, suppose we wish to study the effects of information about an election on voter turnout in that election. We could go out and measure how informed various citizens are about the election through surveys and then later measure the extent to which these individuals participate in the election. If we find that the more informed voters participate more, we might conclude that being informed is motivating these voters to turn out. However, such a conclusion would be premature. An alternative interpretation of the correlation between information and participation could be that citizens who intend to vote are more likely to seek out information. Perhaps the intention to vote has a causal effect on the demand for political information, and therefore also on knowledge. If the demand for information is a function of voting intentions, a statistical correlation between information and participation is not evidence that being informed has a causal effect on the likelihood of voting. Since choosing to be informed is a function of the outcome variable under investigation (turnout), then even if the researcher finds more informed people to be more likely to participate, it should not be concluded that information is causing participation. If uninformed voters were informed, they would not participate—and if the informed voters were uninformed, they would still participate. We have an endogeneity problem.

Endogeneity—Why Is It Particularly Relevant For Public Management Research?

These two sources of endogeneity are particularly relevant to public management research. Selection problems and simultaneity problems likely occur whenever assignment to the independent variable can be chosen (as in the example of military service above) and particularly when this choice is likely to be systematically related to the dependent variable. These problems are particularly relevant whenever the independent variable can be manipulated by policy makers and civil servants. Policy makers and civil servants react to problems and try to solve them by means which they think—or hope—will work. Problems thus lead to solutions, which again influence the problems. If the policy makers or civil servants are correct, there is a loop of causality between the independent and dependent variables, or, in other words, simultaneity. In this sense, public management research is a field where some actors deliberately seek to create endogeneity problems for research.

It is not difficult to think of practical examples. Well-known reforms within public management, such as contracting out, reorganization, decentralization, devolution, delegation, municipal amalgamations, budgetary reforms, privatization and performance management, are rarely introduced without reason. The basis for most reforms is typically a performance problem. Indeed, it is often difficult to justify their introduction otherwise. In that sense, the reforms are endogenous, a fact which creates problems for researchers attempting to measure their effects on performance (Besley and Case 2000).

In light of this challenge, it is lamentable that so few public management researchers turn to experimental methods. The lack of interest is not due to a lack of encouragement. Twenty years ago, for example, the *Journal of Public Administration Research and Theory* wanted to support experimental studies and began a series on experimental designs in public management research (Bozeman 1992). After a handful of articles, however, the supply chain dried up, and the series was stopped. In a recent overview of the state of experimental methods in public management research, Margetts (2011) concluded that little had changed since the *JPART* initiative. The increasing use of experiments in other social science areas in recent decades had largely gone unnoticed within public management research. This lacuna is one of the consequences of the separation of public management research from mainstream social science (Kelman 2007, 226-7).

We strongly believe that this state of affairs ought to be changed and that now there is a great potential in doing so. We do not wish to discount the challenges facing experimental studies, and we recognize that public management researchers may even face special challenges (Bozeman and Scott 1992). But we want to point out that a two-pronged case can be made for more experimental studies in public management research. Not only are the endogeneity problems leading researchers in other areas of the social sciences to experimental methods especially severe within public management research, this field also offers a number of special opportunities for conducting experimental studies. Policy makers and civil servants are often genuinely interested in evidence-based solutions. They are therefore often interested in cooperating with public

management researchers about experiments (Stoker and John 2009; Stoker 2010). Although cooperation between practitioners and researchers is far from straightforward, notable results have recently come out of such cooperative endeavors (see, e.g., Grimmelikhuijsen and Meijer 2012; Haynes et al. 2013; Jakobsen 2013; Bellé 2013; Dynarski et al. 2013).

Statistical Or Design-Based Solutions?

In public management research we cannot assume that an empirical correlation between x and y can readily be interpreted as an estimate of the causal effect of x on y. In fact, as we saw above, we can often (whenever simultaneity is a problem) be pretty sure that the relationship between x and y goes both ways. Only rarely can we rule out, without any reasonable doubt, that changes in x could possibly be introduced in order to solve some problem related to y. Whenever that may be a possibility, simultaneity is an issue to be taken seriously.

The problem can be addressed in two ways: one statistical, the other related to research design. The statistical approach is briefly presented before we discuss the design-based approach in greater detail. The next section introduces five important experimental designs that all go to the root of the endogeneity problem.

A bivariate correlation between an independent (x) and dependent variable (y) is renowned for possibly being biased. Bias occurs, for instance, whenever y is affected by some other variable than x and when this variable is correlated with x. This problem can

be addressed by statistical control. This approach is what the workhorse of statistics, OLS multiple regression, is all about. This solves problems of endogeneity caused by selection bias and omitted variables, but only if relevant control variables can be identified and measured precisely. When the root of endogeneity is simultaneity, statistical control is of little help: If y has a causal effect on x, this is something that we cannot control for. We cannot control for the dependent variable in a statistical analysis of the determinants of the dependent variable. In other words, simultaneity lames the workhorse; statistical control will not help.

In order to find a statistical remedy for the simultaneity problem, we must find a way of getting hold of exactly that part of x that has a causal effect on y but that is not due to y. This can be attempted by the instrumental variable (IV) approach. The method is intuitively simple: Find some variable z, which is correlated with x (preferably strongly), but which we for some reason know cannot be affected by y. Then estimate x from z and use predicted values for x as the independent variable in a regression analysis of y. Since we know that y cannot affect z, we also know that the statistical correlation cannot be due to reverse causality. We refrain from introducing the IV approach in detail (introductions can be found in most introductions to econometrics, see also Angrist and Krueger 2001). Instead, we provide an example and discuss some of the most important limitations of this approach.

Hoxby (2000) is interested in the effect of competition on student achievement in public schools. She faces a serious endogeneity problem, since variation in competition (the

supply of school districts) is likely to be determined in part by factors related to student achievement (Hoxby 2000, 1214). She solves the problem by finding an IV, that is, a variable that affects competition but cannot be correlated with factors associated with the dependent variable. She uses the number of streams within an area, since, "the number of school districts in a given land area at a given time of settlement was an increasing function of the number of natural barriers. I focus on streams, because they are the most common and most easily quantified natural barriers" (2000, 1216). She proceeds to show how streams predict competition well, making a convincing argument (which, of course, cannot be established empirically with certainty) that streams are not caused by school productivity. She finds that competition does in fact improve performance.

However, the statistical approach to the endogeneity problem has its limits. First and foremost, good IV's can be hard to find. Second, no statistical procedure can settle the question of whether an IV is not, directly or indirectly, affected by the dependent variable. This problem is not trivial, since just a feeble effect of the dependent variable on the IV can imply significant bias, especially when the correlation between the IV and dependent variable is small (Bound et al. 1995). In other words, the statistical approach is only sometimes available; and when it is, no methods can prove that the IV actually solves the problem. IV estimation is a very useful tool, but it is not always the optimal solution.

Experiments As A Design-Based Solution

The logic of the design-based approach is to eliminate the source of the endogeneity problem. If the researcher ensures, when collecting data, that the independent variable is exogenously determined, it becomes possible to rule out entirely and with certainty that a possible empirical correlation is due to endogeneity (Dunning 2012, 24–25). An experiment is a special test. In public management studies, a researcher often wants to identify the effects of some initiative, such as a reform or organizational change. In the following, we call this initiative the experimental intervention. Further, an experiment typically has one group (the experimental group) which is subjected to the intervention, while another (the control or baseline group) is not. The effect of the intervention is then identified as the development in the experimental group compared to the baseline group. Concepts like intervention, experimental group, control or baseline group and comparison are thus central in experimental research.

However, defining an experiment more precisely is not easy. The experiment literature disagrees on the exact definition. Morton and Williams (2010, 42) define an experiment as a test where the researcher controls the intervention and actively manipulates it. Dunning (2012, 15–16) adds that the assignment to experimental and control or baseline groups must be random. Randomization is also central for McDermott (2013, 608). Sekhon and Titiunik (2012, 35) also focus on randomized assignment to experimental and control groups but do not consider the researcher's control over the intervention part of the definition of an experiment. The conceptual challenge is that there are many different types of experiments that a definition must encompass. But none of the mentioned definitions covers all forms of experiments. We therefore prefer to use a

less fine-grained definition of the basic concept of an experiment before defining the different forms of experiments more precisely.

For a basic definition, we return to Cook and Campbell's (1979, 5) classic introduction to experimental analysis. Their definition of an experiment is this: "All experiments involve at least a treatment, an outcome measure, units of assignment, and some comparison from which change can be inferred and hopefully attributed to the treatment". This definition includes an external intervention but not the researcher's control over it or over the environment. It also includes assignment to experimental and control groups, but assignment is not necessarily randomized. The elements that are not included in the definition can be used to distinguish between different forms of experiments. We make these distinctions in Table 1.

Table 1 draws a distinction among different forms of experiments. We discuss each form in greater detail in the next section. But first we focus on the five criteria used to categorize the different experiments. The first criterion is whether comparison between experimental and control groups or another baseline is made. Evidently, all forms of experiments make such comparisons. This characteristic is thus generic to all experiments. Other research methods may also work with experimental and control groups, however, such as large-N observational studies. This criterion is consequently not useful for distinguishing experiments from other research methods.

The second criterion is whether the experimental intervention is exogenous. This criterion also applies to all types of experiments. Moreover, it distinguishes experiments from other research methods. There is no guarantee of exogeneity in non-experimental designs, neither in large-N designs nor in single case studies. Endogeneity and exogeneity are rarely absolute concepts; both are normally present to some extent. If a researcher can argue that although an intervention is not completely exogenous, it is not influenced by the units under study to such an extent that it influences the effect measures, then we may speak of "as-if" exogeneity.

The third criterion is whether assignment to experimental and control groups is randomized. Randomization need not be done by the researcher. Sometimes, natural events or political decisions affect different parts of the population or the country in a random manner. In this case, we speak of "as-if" randomization (Dunning 2012, 9–10). If there is no randomization, different results for experimental and control groups may be caused by initial differences between these groups. In this case, the researcher must deal with these differences before interpreting the results. In contrast, randomized assignment means that the causal effect of the intervention may be directly measured as the difference in the result for the experimental and control groups. As evident in Table 1, randomization is not a generic characteristic of experiments, since quasi-experiments do not randomize assignment to experimental and control groups. Furthermore, randomization is not always implemented perfectly in field experiments, as subjects may drop out of an experiment, it may be difficult to prevent communication between subjects

in experimental and control groups, or subjects may fail to respond when post-experiment surveys are conducted (Green 2010; Sinclair et al. 2012).

The fourth criterion is whether the researcher controls the intervention. If so, the experimental variation can be fine-tuned more precisely to the research question than in the opposite case, where the researcher can only analyze the effect of the variation provided by nature or the political-administrative system. Control of the intervention is not a generic characteristic of experiments, but something that characterizes certain experiments.

Finally, the fifth criterion is whether the researcher controls the environment. Control of the environment can be used to minimize or eliminate disturbances from the outside of the intervention and/or control group. Again, this is not a generic criterion. Complete control of the environment is only possible to achieve in the lab where subjects can be isolated during the experiment.

In sum, the central difference between experimental and non-experimental research is not whether experimental and control groups are used, whether assignment to groups is randomized, or whether the researcher controls the experimental intervention. The distinguishing characteristic is the exogeneity of the intervention. All experiments have exogenous interventions. There is no guarantee of exogeneity in other research designs. Therefore, all experimental designs effectively address endogeneity problems due to reverse causality.

EXPERIMENTAL DESIGNS

The five types of experiments outlined in Table 1 are discussed in greater detail below. We present a precise definition, discuss pros and cons and provide an illustrative example. We also provide references on where to find thorough introductions to each of the five types.

The Lab Experiment

Most laboratory experiments[2] exemplify the characteristics described above: exogenous random assignment of treatments to subjects, a baseline group for comparison, and, importantly, control over other factors that may influence the outcome which are not easily controlled outside the laboratory. The laboratory is particularly useful as a way of comparing different institutions, a central issue in public management. That is, usually when we compare different governmental institutions across political jurisdictions with observational data, we must control for differences in culture, preferences and other factors that vary across the jurisdictions. In the lab, however, we can control for these variations and isolate the effects of the variation in institutions alone.

Consider the comparison of different voting mechanisms—e.g. methods of determining the winner in a three-candidate election. In some jurisdictions, no matter how many candidates are running, the winner is the candidate with the most votes (referred to as "first past the post" elections or plurality rule). In other jurisdictions, however, a candidate is required to receive a majority (more than 50% of the votes) to be elected the

winner; a run-off election is held between the two top vote winners if no candidate initially receives a majority.

In a three-candidate election, plurality rule may theoretically lead to a candidate winning office who has received less than a majority of the vote. In such cases, that candidate may actually be what is called a "Condorcet loser"; that is, he or she would have lost the election if running against just one of the other candidates. Specifically, suppose that there are three candidates, A, B and C. Thirty percent of the electorate prefer A to B to C, thirty percent prefer B to A to C, and 40 percent prefer C to either A or B (they are indifferent between A and B). If everyone votes for their most preferred candidate (their sincere preference), and the election is decided by plurality rule, then C will win, even though if the election were just between C and A or C and B, C would lose. A and B supporters would like to coordinate on one of those candidates (by, say, A supporters voting strategically for B and B voters voting sincerely for B) to defeat C. Doing so is an equilibrium under plurality rule but requires some mechanism by which the A and B supporters coordinate on B as their common choice.

In contrast, majority requirements do not require this type of coordination. If everyone votes sincerely, then C receives the most votes, but does not win outright and must face either A or B in a run-off election; where C will be defeated. Thus, majority requirements theoretically function as an institutional coordination mechanism to prevent Condorcet losers from being elected.

Since political jurisdictions vary over whether they use plurality rule or majority requirements (e.g., majority requirements are used in some states and cities in the United States and France, whereas they are not used in other states and cities in the US or in the UK), it would seem as though we could merely compare how these systems work in these various jurisdictions to see if the theoretical predictions are correct. But there are many other differences across these jurisdictions that also vary, including the number of candidates, voter preferences, the ability to coordinate across different groups, voter information levels and so forth. Moreover, majority requirements might specifically be used in jurisdictions where coordination is more difficult and there are likely to be more than two candidates, whereas plurality rule is used in jurisdictions where coordination is easy and/or there are typically only two candidates (and thus no Condorcet loser problem). Thus, we may also have an endogeneity problem.

In the laboratory, we can investigate the two types of voting mechanisms holding the number of candidates, information, coordination abilities and voter preferences constant. We can induce voter preferences by paying subjects based on the winner of the election. So, for example, we can assign three voters to receive payoffs of, say, $1 if A wins, $0.75 if B wins and $0.25 if C wins; three voters to receive payoffs of $1 if B wins, $0.75 is C wins and $0.25 if C wins; and four voters to receive payoffs of $1 if C wins and $0.25 if either A or B wins. Keeping these numbers of voters and payoffs constant, we can then compare how voters choose under plurality rule versus elections with majority requirements. Morton and Rietz (2008) conduct such an experiment. They demonstrate that, as theoretically predicted, coordination of A and B voters is extremely difficult

under plurality rule and the Condorcet loser can often win. Furthermore, they show that majority requirements do indeed significantly reduce the possibility of the Condorcet loser winning.

This example is a lab experiment evaluating the effects of different institutions that exist in the naturally occurring world but are difficult to investigate with observational data due to the endogeneity problem and the inability to control for the many possible confounding factors. Importantly, however, the laboratory can go beyond the study of existing institutions to investigate proposed institutions that have never been used or ones that are used rarely. For instance, Gerber et al. (1998) investigate a voting system, which has been advanced as a method of increasing minority representation: cumulative voting. In cumulative voting, voters are allowed more than one vote, which they can cast either in bulk for one candidate or distribute across multiple candidates. The method was used in the state of Illinois in the 19th century and is used in just a few small localities in the United States. Thus, there is insufficient naturally occurring data to evaluate this mechanism. By using lab experiments, Gerber et al. (1998) were able to study the institution without either waiting for a government to adopt that institution or having to convince government officials to consider adopting the institution even as a field experiment, something that can be extremely difficult given that the institution might affect real policy outcomes. Furthermore, lab experiments enable measurement of variables that can otherwise be very difficult to track. For example, preferences can be measured, or induced, precisely in the lab. Outside of the lab this can typically only be done by cruder methods. Other examples could be eye tracking or measurement

physiological parameters such as breathing and muscle activity. Such information is nearly impossible to collect outside of the lab.

Obviously, lab experiments have limitations. In the laboratory, we can typically work with only small subject pools at a time; Morton and Rietz' experiments used 28 subjects in each session, which is much smaller than is typical in most elections in which either plurality rule or majority requirements are used. Furthermore, the environment may seem artificial to subjects and they might therefore behave differently in the lab than they would in an election outside the laboratory. One of the main reasons for using financial incentives is to induce subjects to take the laboratory environment seriously and to have as much at stake in the laboratory election as they would in a naturally occurring election, thereby mitigating the effects of artificiality. Lab experiments also often use student subjects, who may not be representative of the population at large. It is of course possible to conduct lab experiments with nonstudents or more representative samples, but typically more costly.

The Survey Experiment

The survey experiment[3] is also a pure experimental design: The researcher controls an exogenous intervention in a treatment group, and the assignment of subjects to the treatment and control group is randomized.

The survey experiment stands out because of its delivery vehicle. They are conducted in surveys by asking different versions of survey questions to groups of respondents. The

respondents are randomly assigned to groups. Hence, the intervention is the difference in how the survey question is phrased, and, due to randomization, any difference in answers can be interpreted as an effect of the treatment.

Nielsen and Baekgaard (forthcoming) are interested in the effect of performance information on politicians' spending preferences. This question is important, because although performance information systems have become widespread, we know little about how this information affects decision-makers. Nielsen and Baekgaard argue that information performance information indicating low and high performance in some services has a positive effect on preferences for the spending on these services, while average performance information does not affect spending preferences. These claims are difficult to investigate empirically with traditional observational data due to two methodological problems. First, performance information may be correlated with, for example, the ideological inclinations of the majority, which may again be correlated with both knowledge of performance and spending preferences. Second, simultaneity is a serious issue if spending preferences affect how performance information is gathered (2013, 17). The authors solve this problem by using a survey experiment. In the survey, 844 local councilors were randomly divided into four groups. Respondents in each group were treated with a special information cue. In the control group, no information on performance was provided. In three treatment groups, information indicating low, average and high performance was provided. All respondent were then questioned about their spending preferences. The random assignment of respondents ensures that any differences in spending preferences between the groups must be due to performance

information. The authors find that performance information does in fact matter; politicians become more inclined to prefer higher spending if they have been told that performance is not neutral.

One of the most attractive features of the survey experiment is that it is relatively cheap and easy to implement. The costs are typically comparable with those of an ordinary survey, and a single survey can easily accommodate multiple experiments. Another advantage is that survey experiments allow for a large number of respondents. Hence, this design is compatible with standard large-N methods, estimations of moderating effects and so on. A limitation of the survey experimental design is that they typically cannot be incentivized and that the intervention is of low intensity. Compared to lab and field experiments, where subjects can be affected economically, socially or materially by the treatment, subtle changes in the wording of survey questions are quite weak. Furthermore, as other surveys, survey experiments can suffer from considerable nonresponse, which limits the validity of the results as estimates of true population parameters.

The Field Experiment

The field experiment[4] is closely related to the lab experiment. As in the lab experiment, control and experiment groups are compared, the intervention is exogenous, assignment to the treatment or control group is randomized and the researcher actively controls the intervention.

Not surprisingly, the difference is that while a lab experiment proceeds in a controlled environment, the field experiment takes place in real world settings, where interventions, subjects, context and outcome measures reflect what actually happens (Davenport et al. 2010, 69–71). Hence, if studying a hypothesis regarding decision-making procedures in the European Union (EU), the field will be the EU institutions, and a true field experiment would (typically unrealistically) involve a randomized intervention in actual decision-making rules. If the hypothesis concerns the effect of voter registration on turnout, a field experiment would involve randomized changes being made in how voters are registered (as in Gosnell, 1927, which according to Davenport et al. (2010) was the first field experiment in political science).

Transformational leadership may be a powerful tool when public managers wish to motivate employees. It is, however, notoriously difficult to study in ordinary observational studies. The potential simultaneity problem is apparent: Leaders are likely to react to employees' motivations by adjusting their leadership strategies. Perhaps transformational leadership tends to be employed when motivation is low. Or maybe leaders use this strategy mostly when motivation is already high. The correlation between the use of transformational leadership and motivation therefore says little concerning the causal effect of using this strategy.

Bellé (2013) addresses this problem by using a field experiment. He expects transformational leadership to improve performance but that the effect is moderated by factors such as whether employees are in close contact with those benefitting from the

service. In the field experiment, 138 nurses at a public hospital were randomly divided into six groups. The first group (baseline group) was shown a tutorial video on how to perform a specific task. Subjects in a second group were shown the same video, together with a transformational talk by the Director of Nursing. In the third group, the video was shown and subjects met a patient that had benefitted from the specific task, while members of the fourth group were exposed to the video, the talk and the patient.[5] After the intervention, the nurses completed the tasks and performance was measured. The effect of the transformational leadership talk can then be estimated by comparing the respective performances of the subjects in groups 1 and 2 and of meeting the patient by comparing 1 and 3. The moderating effect of contact with beneficiaries can be estimated by comparing the performance in group 4 with groups 2 and 3. These comparisons allow Bellé to conclude that transformational leadership can have a major impact, particularly when employees are in contact with beneficiaries.

The randomized assignment of nurses to the six groups ensures that differences in performance between the groups cannot be due to self-selection,the public managers' reactions to prior performance or any other irrelevant factors, provided, of course, no drop-out or nonresponse. The performance effects must be due to the intervention.

Like other experimental designs, a field experiment ensures a high degree of internal validity. Contrary to other experimental designs, however, field experiments can also have quite high ecological validity; that is, since they take place within a naturally occurring environment, the treatments that subjects experience are arguably more "real"

and seem more natural than those in the laboratory.[6] Field experiments can also be highly relevant for policy-making. Field experiments in public management are typically only possible if some public organization considers reform. If the reform is designed as a field experiment, the practical relevance of the study is almost guaranteed. The results will tell exactly what policy-makers are (or should be) interested in: whether the reform works as intended.

Field experiments also have some obvious drawbacks. First, they are often relatively expensive to carry out, as they require acceptance from a number of actors from outside of academia, and real world interventions often require resources. Second, many important research questions are difficult to address by field experiments. This is the case, for example, when formal rules regulate who is supposed to be subject to the intervention. In such cases, legal considerations, not randomization, will determine allocation to the treatment and control groups. Third, politically salient topics are hard to study in field experiments. Randomization can then be difficult to defend politically. For example, a field experiment on the effect of extra teacher resources in the classroom can be hard to randomize, because politicians will have to defend why these resources are not allocated to, for instance, poorly performing schools. Fourth, field experiments typically require substantial coordination with public authorities or citizens, who can be hard to control and possibly care little about the scientific validity of a research design. Fifth, field experiments are harder to control than other experimental designs. For instance, it is often impossible to ensure that participants in the control and intervention groups do not communicate with each other. This communication makes it harder to distinguish clearly

between the two groups. Finally, field experiments which may put some citizens in a worse position can be problematic from an ethical perspective.

The Natural Experiment

Natural experiments[7] are so named because their data stem from nature alone without the intervention of a researcher. Within the social sciences, however, the data for natural experiments are normally products of the political-administrative process. The defining characteristic is that the data are provided from the outside, not from manipulations by the researcher. In this sense, natural experiments are different from lab, field and survey experiments. But like these experiments, natural experiments include randomized (or "as-if" randomized) experimental and control groups. Further, natural experiments share with other experiments the characteristic that the experimental intervention is exogenous (Dunning 2012, 41–63).

Compared to other research designs, natural experiments possess advantages and disadvantages. Compared to traditional observational studies—both large-N studies and single case studies—natural experiments address endogeneity problems more effectively, since the experimental intervention is exogenous. Natural experiments also address the problem of omitted variables more adequately, since the assignment to experimental and control groups is randomized or "as-if" randomized. Compared to other experimental designs, natural experiments possibly investigate the effect of factors that are difficult to manipulate by the researcher, even in a lab. For example, how does military conscription affect the political attitudes of men? Do police patrols influence crime rates? Do election

observers have an effect on election fraud? These are questions that have been examined by natural experiments (Erikson and Stoker 2011; Di Tella and Shargrodsky 2004; Hyde 2007) and are difficult to analyze by lab experiments. They are also difficult to analyze by traditional observational studies due to endogeneity problems.

The drawback of natural experiments is the fact that the researcher depends on outside forces for data, since they are created either by nature or political initiatives. The lack of control introduces some randomness as to the exact questions that may be investigated by natural experiments. It also means that the researcher can only analyze the variation offered by outside forces. The experimental intervention can rarely be fine-tuned to provide the optimal test of the researcher's hypothesis.

One example of a natural experiment is Angrist et al.'s (2002) study of private school vouchers. The researchers wanted to investigate whether providing children with vouchers for private schools improves their educational performance. They exploited the fact that the government in Colombia had established a program offering such vouchers to children from low-income families. The vouchers were distributed on a neighborhood basis, and in cases where demand exceeded supply, lotteries were used to allocate the vouchers among eligible children who wanted to participate in the program. To the researchers, this set-up represented an experiment with randomized assignment to experimental and control groups. They used the government program to investigate the effects of vouchers on different measures of school performance. Data were collected by surveys to program applicants; that is, to both winners and losers in the lottery. The

experimental intervention was exogenous to the individual student's prior educational performance and interest in being enrolled in the program, and assignment to experimental and control groups was randomized. The analysis therefore effectively dealt with endogeneity problems. Further, the experiment was a natural one because the experimental intervention was provided from the outside: the Colombian government. This lack of control over the intervention meant that the researchers could not manipulate exactly according to their theoretical interests—for example to examine the effect of varying the monetary value of the voucher or varying the program's target groups.

The Quasi-Experiment

Quasi-experiments[8] resemble other experiments in the sense that there are experimental and control groups, an exogenous intervention and measures of the effect of the intervention. Further, quasi- experiments share the characteristic with natural experiments that the intervention comes from the outside. It is not controlled and manipulated by the researcher, but created by nature or the political-administrative system. In contrast to other experiments, however, assignment to experimental and control groups is not randomized in quasi-experiments. The two groups may therefore differ in other respects than their exposure to the experimental intervention. Quasi-experimenters therefore face the challenge of interpreting whether differences in results are caused by the experimental intervention or by initial differences between the groups (Shadish et al. 2002).

Quasi-experiments possess distinct advantages and disadvantages. Compared to traditional observational studies, they address endogeneity problems more

effectively, because the experimental intervention is exogenous. But within the group of experimental designs, it is the type of experiment that provides the researcher with the least leverage over the research question. This lack of influence arises because assignment to experimental and control groups is not randomized. In contrast to other experimental designs, quasi-experimenters must therefore deal with initial differences between groups. This challenge is not small, since it can rarely be ruled out that unobserved differences and some endogeneity problems do not remain even after the careful collection of data for control variables.

Yet quasi-experiments possess considerable potential. They are stronger designs than traditional observational studies, since they tackle endogeneity problems more directly. Further, quasi-experiments, like natural experiments, render it possible to investigate the effects of phenomena that are difficult to manipulate by the researcher. Consequently, if randomization can never be sacrificed, many research questions must remain unanswered.

An example of a quasi-experiment is Blom-Hansen et al.'s (2014) study of scale effects in local government. The researchers wanted to investigate whether the economic costs of governing are influenced by jurisdiction size. They used a municipal reform in Denmark that led to the amalgamation of a large number of municipalities but which also left a number of municipalities intact. The authors argued that the change in jurisdiction size for the amalgamating municipalities was not caused by local factors but instead by an external intervention from the national government. Concerns of reverse causality were

thus minimized. Assignment to experimental and control groups was not randomized, and there were certain structural differences between the groups. For example, municipalities on islands were difficult to amalgamate. Post-reform efficiency differences might therefore to some extent be caused by initial differences between the groups. The researchers therefore included a number of control variables which minimized concerns of omitted variable bias. Since both reverse causality and omitted variable bias were effectively addressed, the researchers had an "as-if" exogenous intervention. Based on a difference-in-difference design in which *pre*- and *post*-reform efficiency measures were compared for both experimental and control groups the authors concluded that there are positive scale effects of jurisdiction size.

Internal And External Validity In Experimental Designs

The lab experiment with randomized assignment to experimental and control groups, and complete control by the researcher over the experimental intervention is often considered the scientific gold standard. Other experimental designs may only aspire to a silver standard. Our discussion of four other experimental designs—field, survey, natural and quasi-experiments—provides a more nuanced picture. These experimental designs all possess distinct advantages and disadvantages, the most important being that they allow the investigation of research questions that are difficult to study in the lab. Therefore, we welcome not only the growing interest in and increasing use of experiments, but also the use of mixed experimental designs, such as lab-in-the-field experiments (Morton and Williams 2010, 296-301) and Internet-based experiments combining lab and survey experiments (Eckel and Wilson 2006).

Common to all these types of experiments is that they can provide the empirical basis for valid causal inference. They have, in other words, a high degree of internal validity. Some argue that this comes at a cost in the sense that the external validity is then bound to be correspondingly low. Internal and external validity are often seen as each other's opposites. But we think that this approach is simplistic and obscures the debate. Both concepts are multifaceted, and there is no logical trade-off between them (Morton and Williams 2010, 253-276). Internal validity may be decomposed into construct, statistical and causal validity. Construct validity has to do with how closely the data measure the theory, while statistical validity deals with the empirical measurement of the relationship between dependent and independent variables. There is no logical reason why experiments should hold any particular advantage compared to observational studies when dealing with these aspects of internal validity. They are challenges for both types of research. It is only in addressing the third aspect of internal validity, causal validity, that experiments may possess an advantage. Since experiments can ensure that causality only runs in one direction, they are better able to address problems of reverse causality. Since this is a major challenge for public management research, we believe that this is a powerful argument for the use of experiments in this line of research.

Turning to external validity, there is often significant confusion about the strengths of observational studies and weakness of experimental research. External validity refers to how valid results are beyond the setting in which they are studied. External validity is thus different from statistical inference, which is part of a study's internal validity. Even the finest observational study of, say, a representative sample of

public organizations in the United Kingdom, may not be externally valid for China. External validity is partly composed of ecological validity. This aspect of external validity has to do with how similar the setting of the study is to the real world. Observational studies often face fewer problems here since their data are by definition from the real world. Experiments, on the other hand, may face a problem, but this may be the case to different degrees. Lab experiments are by definition conducted in an artificial environment, but field experiments are, also by definition, conducted in the real world. However, ecological validity does not equate to external validity. Indeed, there may be difficult trade-offs involved. Increasing a study's ecological validity by, for example, doing experiments in the field rather than the lab, may decrease the overall external validity of the study since the results may then apply to a lesser degree to settings different from the chosen field location.

CONCLUSION

Our main message in this paper is that experimental research has something to offer public management research. In this line of research, endogeneity in the form of selection problems and simultaneity, or, reverse causality, represents a special challenge. The objects of study – organizational reforms, budget systems, decentralization, and so on – are rarely changed without any thoughts on their effects. Indeed, policy makers and civil servants normally introduce changes exactly because they think that they will work in some sense. So, problems lead to solutions, which again influence the problems. There is a loop of causality between the independent and dependent variables which may be

impossible to disentangle for observational studies. This is why we think experimental studies have something to offer.

However, some may object that there is a price to be paid if public management researchers turn to experimental studies. Many researchers acknowledge that experimental studies are better at addressing causality issues, but argue that although experiments may thus be high on internal validity, they are low on external validity. We argued above that this objection is invalid. Indeed, external validity is a challenge for both experimental and observational research. There is no foolproof way – neither for the observational, nor for the experimental researcher – of establishing how externally valid the results are. The proof must be established gradually over time as studies are replicated in different settings and meta-analyses can be conducted. This is the case for both observational and experimental research.

In sum, there is no logical contrast between internal and external validity. Turning to experimental studies in public management research therefore carries no obvious risk in terms of external validity. Since there is a potentially large benefit to be harvested in terms of more efficient handling of endogeneity problems, the choice appears to be an easy one. There is simply no compelling reason why public management researcher should not do more experiments. This, however, does not mean that all public management researchers should abandon ship and stop doing observational studies. As we have argued in the paper, not everything is amenable to experimental manipulation, so both observational and experimental techniques belong in our methodological tool box.

REFERENCES

Angrist, J., E. Bettinger, E. Bloom, E. King and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment", *American Economic Review* 92(5): 1535–1558.

Angrist, J. and A. B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments", *Journal of Economic Perspectives* 15(4): 69–85.

Bellé, N. 2013. "Leading to Make a Difference: A Field Experiment on the Performance Effects of Transformational Leadership, Perceived Social Impact, and Public Service Motivation", *Journal of Public Administration Research and Theory* (advance access published 13 June 2013).

Besley, T. and A. Case. 2000. "Unnatural Experiments? Estimating the Incidence of Endogenous Policies", *Economic Journal* 110(467): F672–F694.

Blom-Hansen, J., K. Houlberg and S. Serritzlew (2014). "Size, Democracy, and the Economic Costs of Running the Political System", *American Journal of Political Science* 58(4): 790–803.

Bound, J., D. A. Jaeger and R. M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association* 90(430): 443–50.

Bozeman, B. 1992. "Experimental Design in Public Policy and Management Research: Introduction", *Journal of Public Administration Research and Theory* 2(3): 289–292.

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

Cook, Thomas D. and Donald T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Davenport, T. C., A. S. Gerber and D. P. Green. 2010. "Field Experiments and the Study of Political Behavior". Pp. 69-88 in J. E. Leighley, ed., *The Oxford Handbook of American Elections and Political Behavior*. Oxford: Oxford University Press.

Di Tella, R. and E. Schargrodsky. 2004. "Do Police Reduce Crime? Estimates Using the Allocation of Policy Forces after a Terrorist Attack", *American Economic Review* 94(1): 115–133.

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

Dynarski, S., J. Hyman and D. Whitmore Schanzenbach. 2013. "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion", *Journal of Policy Analysis and Management* (early view, article first published online 26 July 2013).

Eckel, C. C. and R. K. Wilson. 2006. "Internet Cautions: Experimental Games with Internet Partners", *Experimental Economics* 9(1): 53–66.

Erikson, R. and L. Stoker. 2011. "Caught in the Draft: The Effects of Vietnam Draft Lottery Status on Political Attitudes", *American Political Science Review* 105(2): 221–237.

Gerber, A. S. 2011. "Field Experiments in Political Science". Pp. 115-141 in J. N. Druckman, D. P. Green, J. H. Kuklinski and A. Lupia, eds., *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press.

Gerber, E., R. Morton and T. Rietz. 1998. "Minority Representation in Multimember Districts," *American Political Science Review* 92(1): 127–144.

Gosnell, H. F. 1927. *Getting-Out-The-Vote: An Experiment in the Stimulation of Voting*. Chicago: University of Chicago Press.

Green, J. 2010. "Points of Intersection between Randomized Experiments and Quasi-Experiments", *Annals of the American Academy of Political and Social Sciences* 628: 97–111.

Grimmelikhuijsen, S. G. and A. J. Meijer. 2012. "The Effects of Transparency on the Perceived Trustworthiness of Government Organization: Evidence from an Online Experiment", *Journal of Public Administration Research and Theory* (advance access published 5 November 2012).

Haynes, L. C., D. P. Green, R. Gallagher, P. John and D. J. Torgerson. 2013. "Collection of Delinquent Fines: An Adaptive Randomized Trial to Assess the Effectiveness of Alternative Text Messages", *Journal of Policy Analysis and Management* (early view, article first published online 6 August 2013)

Hoxby, C. M. (2000). "Does Competition among Public Schools Benefit Students and Taxpayers?", *American Economic Review* 90(5): 1209–1238.

Hyde, S. D. 2007. "The Observer Effect in International Politics: Evidence from a Natural Experiment", *World Politics* 60(1): 37–63.

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

Kelman, S. 2007. "Public Administration and Organization Studies", *The Academy of Management Annals* 1(1): 225-267.

Margetts, H. Z. 2011. "Experiments for Public Management Research", *Public Management Review* 13(2): 189–208.

McDermott, R. 2013. "The Ten Commandments of Experiments", *PS: Political Science and Politics* 46(3): 605–611

Morton, R. and T. Rietz. 2008. "Majority Requirements and Minority Representation", *NYU Annual Survey of American Law* 63(4): 691–726.

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

Nielsen, P. A. and M. Baekgaard. Forthcoming. "Performance Information, Blame Avoidance, and Politicians' Attitudes to Spending and Reform: Evidence from an Experiment", forthcoming in *Journal of Public Administration Research and Theory*.

Sekhon, J. S. and R. Titiunik. 2012. "When Natural Experiments are Neither Natural nor Experiments", *American Political Science Review* 106(1): 35–57.

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

Sinclair, B., M. McConnell and D. P. Green. 2012. "Detecting Spillover Effects: Design and Analysis of Multilevel Experiments", *American Journal of Political Science* 56(4): 1055–1069.

Sniderman, P. M. 2011. “The Logic and Design of the Survey Experiment: An Autobiography of a Methodological Innovation”. Pp. 102-115 in J. N. Druckman, D. P. Green, J. H. Kuklinski and A. Lupia, eds., *Cambridge Handbook of Experimental Political Science*. Cambridge: Cambridge University Press.

Stoker, G. and P. John. 2009. “Design Experiments: Engaging Policy Makers in the Search for Evidence about What Works”, *Political Studies* 57(2): 356–373.

Stoker, G. 2010. “Translating Experiments into Policy”, *Annals of the American Academy of Political and Social Science* 628(1): 47–58.

Wooldridge, J. M. 2013. *Introductory Econometrics. A Modern Approach*. South-Western Cengage Learning.

Table 1. Different forms of experiments

	Comparison of experimental group with baseline?	Intervention exogenous or "as-if" exogenous?	Assignment randomized or "as-if" randomized?	The researcher controls the intervention?	The researcher controls the environment?
Lab experiment	Yes	Yes	Yes	Yes	Yes
Survey experiment	Yes	Yes	Yes	Yes	No
Field experiment	Yes	Yes	Yes	Yes	No
Natural experiment	Yes	Yes	Yes	No	No
Quasi-experiment	Yes	Yes	No	No	No
Traditional large - N observational study	Yes	No	No	No	No
Traditional single case study	No	No	No	No	No

NOTES

[1] Note that this problem can also be seen as an omitted variable problem. The omitted variable problem is another important source of endogeneity. If we could measure motivation and aptitude perfectly, controlling for these variables would remove the correlation between the independent variable and the error term. This control would solve the problem of endogeneity. However, it is not always possible to measure all relevant control variables perfectly (if at all).

[2] See Morton and Williams (2010) for an introduction.

[3] Sniderman (2011) provides an introduction to survey experiments.

[4] Good introductions can be found in Davenport et al. (2010) and Gerber (2011).

[5] Groups 5 and 6 were used to study another potential moderator.

[6] Ecological validity is often confused with external validity, but the two are quite different concepts, cf. our discussion in the conclusion.

[7] Dunning (2012) provides an introduction.

[8] Cook and Campbell (1979) provide a classic introduction.