CEU Political Science Journal. Vol. 6, No. 4

COST AND BENEFITS OF NATURAL EXPERIMENTS IN POLITICAL SCIENCE

Elena Labzina
Central European University Budapest

Abstract
Natural experiment *is a research design that is widely employed in modern Political Science. Yet, in the existing literature the features of the concept are ambiguous. The major aim of this article is to refine the related theory and argue natural experiments provide valid estimates in terms of causal inference. First, the author summarizes briefly and develops the related theory:* the 'as-if' treatment randomization assumption *is redefined with the introduction of* the expected exchangeability treatment assumption*, which enables their classification more as natural experiments. Second, based on the renewed theory, the author proposes an algorithm for the assessment of assumed natural experiments. Third, the algorithm is applied to five illustrative cases of recent natural experiments from Political Science. As a result, it is found that only two of them may be considered capable of providing valid inference. The major empirical finding is these two valid* natural experiments *are 'hidden' experiments, e.g. the individual(s) who were unaware that they assigned treatment had performed the randomization. This leads to the conclusion that the mysterious "nature" in nature experiments is human beings.*

Keywords: assessment algorithm, expected exchangeability condition, 'hidden' experiments, natural experiments

1. Introduction

Natural experiments are in fashion in Social Sciences. According to Google Scholar, during the last five years 3370 articles have been published containing in the name the word-combination "natural experiment", which is almost half of all those published.[1] Yet despite their popularity, the theory on natural experiments is ambiguous. Probably the reason for this is that most of the articles study singular cases while only few investigate the underlying theory and quality of the research on the topic. In particular, it is still not entirely clear which benefits natural experiments bring in terms of causal inference.

Ideologically my paper investigates two almost distinct perspectives, which have been hardly ever analysed together. First, I look at the conditions of the "as-if" randomization and analyse whether they are satisfied in five recent works from

1 Google Scholar: "Natural experiments", search only in Social Sciences, Arts, and Humanities: the overall number of results is 7760, for the period 2006-2010 – 3370. The request was made on September 6, 2011.

Political Science. In a way this part follows Thad Dunning's path.[2] In his paper he evaluates the quality of the natural experiments in terms of the plausibility of the "as-if" randomization assumption. In contrast to his work, I not only analyse the cases, but also develop the theory, making the satisfying of the crucial "as-if" randomization condition more feasible in real-life observational studies relaxing its requirements.

In his work Dunning mentions that other reasons violating "the success of natural experiments" exist, but does not provide any further discussion or assessment. Therefore a "beyond randomization" investigation is the second perspective of my paper, in which I employ the ideas that Jasjeet Sekhon and Rocio Titiunik[3] address in their paper. They underline that even given a perfect randomization, some treatment-control comparisons may not be justified, and even a proper control-treatment contrast may not be related to the investigated effect.

In the subsection on the validities (2.1) to make the overall theory more consistent, I combine the ideas coming from different authors. I find more dimensions of validity to make the theory complete. *Statistical* and *construct validity* are described following Shadish *et al.,*[4]; *ecological validity* is introduced using the ideas of Roe and Lust[5]; to define *content validity*, the understandings of Haynes *et al.*[6], are incorporated. The subsections related to *experiments* (2.2) and *nature* (2.3) present a unique complex view on these seemingly obvious notions. I show that the definition of an experiment may be understood as having four levels and nature may be of three distinct types.

As a result, the overall case assessment in this paper is more detailed than those of Thad Dunning[7] and Jasjeet S. Sekhon and Rocio Titiunik[8]: their ideas along with the

2 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", *Political Research Quarterly* (Oct 2008), 282-293.

3 Jasjeet S. Sekhon and Rocio Titiunik, "When Natural Experiments Are Neither Natural Nor Experiments: Lessons from the Use of Redistricting to Estimate the Personal Vote," *Working Paper* (2010).

4 William R. Shadish et al, *Experimental and Quasi- Experimental Designs for Generalized Causal Inference*, (Houghton Mifflin Company, 2002), 33-103.

5 Brien E. Roe and David R. Just, "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data," *American Journal of Agricultural Economics* (Dec 2009), 1266-1271.

6 Stephen N. Haynes et al, "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods", *Psychological Assessment 7(3)*, (Sep 1995), 239.

7 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 282-293.

refined theory are put together, developed, and "moulded together" into the assessment algorithm. To sum up, this work answers the broad question "do natural experiments provide valid estimates in Political Science?"

Structurally the paper consists of two main parts, theoretical and empirical, and their "mediator", the assessment algorithm. In the theoretical part, I refine and improve the theory to understand the question better and make the answer more precise. First, I summarize the existing theory of validity. Second, I look at the definition of "an experiment" and discover that four levels of understanding exist. Third, I typify "nature" to clarify what "the agent" controlling the experiment is in a natural experiment. Fourth, I come to the crucial condition of an experiment, the treatment randomization. Then the theoretical findings are condensed in the assessment algorithm. In the empirical part, I apply the assessment algorithm to five examples of natural experiments from modern Political Science. The examples and the names of the cases are taken from Dunning's work.[9] Lastly, the findings from both parts are summarized.

The practical purpose of this paper is twofold. On the one hand, it provides a modern theoretical tool to assess a supposed natural experiment. This evaluation may be performed even before the main part of the research, helping to prevent crucial pitfalls to appear in the study. On the other hand, the work presents the actual examples of caveats in modern works from Political Science, which may be generalized for a wider set of researches. Both qualitative and quantitative scientists can make use of this work: the former may be more focused on checking external validity, while the latter on the internal. Nowadays there is a tendency to converge quantitative and qualitative methods, and this paper may be considered a rare example of both.

2. Theory Refinement

2.1 Types of Validity

The purpose of this subsection is to identify the terms in which the quality of a natural experiment can be measured. For this purpose I will clarify the understanding of "validity" and look at its different aspects within "types of validity".

Generally speaking, a valid statement is simply a correct or true statement. But what

8 Jasjeet S. Sekhon and Rocio Titiunik, "When Natural Experiments Are Neither Natural Nor Experiments: Lessons from the Use of Redistricting to Estimate the Personal Vote."

9 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 283.

does "a statement is valid" mean? In terms of causal inference, *the likelihood that an inference is correct* is *the validity of the inference*. In terms of effect estimation, *the extent to which an estimate of an effect is precise* is *the validity of the estimate*. A research method, being a measuring instrument, is valid if it provides valid estimates. To be more specific, it "measures what it is intended to measure"[10] and measures precisely.

In practical terms, validity is not a characteristic of a method but an inference, e.g. an instance of the method's application,[11] since in each application obstacles to the validity are different. Theoretically *the expected validity of a method's application* is approximately equal to *the validity of a method*. Importantly, the validity of a method can be considered multidimensional, where types of validity are different aspects of validity.

The two key types of validity are *internal* and *external* validities (see e.g. the article of Roe and Lust[12]). In the discussion the aim is to maximize their hypothetical sum. More types of validity – *construct*, *statistical*, *ecological*, and *content* - are brought in to assist the evaluation of the major types. While in the literature the definitions of external and internal validities do not differ significantly, the definitions of additional types and even the types themselves may vary. Consequently, it is important to underline where the definitions (for the additional types) employed in the paper come from. *Statistical* and *construct validity* are described following Shadish *et al.,*[13] (2002, 33-103); *ecological validity* is introduced using the ideas of Roe and Lust[14]; to define *content validity*, the understandings of Haynes *et al.*[15], are incorporated.

To start with the key validities, *internal validity* can be defined as the extent to which a proposed causal relation is correct within a research design. More precisely,

10 Edward G. Carmines and Richard A. Zeller, "*Reliability and validity assessment*," (SAGE University Paper, 1979), 17.

11 William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi- Experimental Designs for Generalized Causal Inference,* 34.

12 Brien E. Roe and David R. Just, "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data," *American Journal of Agricultural Economics* (Dec 2009), 1267.

13 William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi- Experimental Designs for Generalized Causal Inference,* 33-103.

14 Brien E. Roe and David R. Just, "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data," 3.

15 Haynes, Stephen N., David C. S. Richard and Edward S. Kubany, "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods", *Psychological Assessment 7(3)*, (Sep 1995), 239.

it is the validity of the statement that the treatment causes the investigated effect in the sample given in the settings. *External validity* is the extent to which a proposed relation defined for an individual design is valid for a more general design. In terms of units, a design has a high external validity if the results obtained based on a sample are valid for the population of interest. In the literature the common view is a trade-off between internal and external validities, and *natural experiments* lie in the middle of the scale.[16]

Among the auxiliary types of validity only statistical validity is directly related to internal validity, while others could be considered to be attributes of external validity. *Statistical validity* assessment is based on answers to two questions. First, do a presumed effect and cause covary? Second, what is the magnitude of the covariation? A significant mathematical covariation as such does not prove a causal inference, but its existence may be a signal to start a deeper qualitative and quantitative investigation. Practically, statistical validity is a technical and mathematical aspect of internal validity. Given an appropriate means of measurement, internal validity causes statistical validity but not *vice versa*.

Content, *construct* and *ecological* validities represent different perspectives of external validity. Assessment of external validity is not restricted to these, but looking at them explicitly may simplify and structure the assessment significantly. *Construct validity* is the validity of the terms used in a research design. Terms or names point to the definitions, or understandings, used in the design. *Ecological validity* is the extent of the possibility to generalize the setting(s) of the design to the universe of all settings of interest. In other words, it is a representation of a "sample" of settings to a "population" of settings for which an inference is performed. *Content validity* is the validity of measuring instruments employed in a design. This can be understood as a part of construct validity, since a measuring approach is embedded, explicitly or implicitly, in the definition of a term.

2.2 What is "an experiment"?

Experiment is one of the terms probably most frequently used in almost all sciences, including Political Science. However, it may often not be completely clear what lies behind the notion, which makes their validity analysis complicated. Hence, before moving further, it must be strictly specified. As an intuitive concept, it is problematic to define shortly and precisely. I argue that from the perspective of Social Sciences an *experiment* can be considered to have four nested definitions, each of which

16 For example, see Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 282-293 or Brien E. Roe and David R. Just, "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data," 1266-1271 .

gradually adds more details to the understanding.

The *basic definition* is intuitive: as Shadish *et al.,* argue, the act of an experiment is "*[t]o explore the effects of manipulating a variable*"[17]. This view includes minimal details and is seemingly too broad to clarify what an experiment is in a more scientific understanding. Contrary to the first one, *the minimal definition* narrows the concept extensively emphasizing that the experiment is a design where *"the treatment [is] randomly allocated over the sample of experiment and is controlled*"[18]. This definition leaves a number of questions unanswered: What is also in the experiment except the treatment? Who assigns the treatment? What is known about the sample? These questions are crucially important from the perspective of internal validity, but remain unclear.

The intermediate definition is probably most often employed in Political Science and attempts to clarify the ambiguity. It states, first, that the effect of the treatment is compared to the effect of no-treatment units called *controls* or *control*. Second, the assignment of the treatments is randomized, which is a version of the minimal definition. Third, the treatment allocation, in addition to other experimental manipulations, must be under the control of the one who performs the experiment.[19]

A few practical remarks need to be made about the latter definition. First, why does it presuppose the presence of an actor who performs the experiment? I argue that this implicitly underlines the need for an actual randomness of the treatment allocation. This means that other sources of the effect are expectedly ruled out. In the case of Political Science, an assumption of the actual randomness - especially while working with historical data - may be excessively strong, and has to be examined carefully for each case. Second, the assumption of the existence of controls as a consequence of the presence of treatment units is extremely tentative. The supposedly 'no-effect' units may turn out to be affected by other "treatments" overlooked at first, since those treatments are unknown, and so not taken into account in the research.

The full definition requires an addition relative to the intermediate one: the sample of an experiment is supposed to be representative of the population for which the

17 William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi- Experimental Designs for Generalized Causal Inference,* 507.

18 Thad Dunning, "Natural Experiments," *Draft entry for the International Encyclopedia of Political Science,* (2010), 2.

19 Based on Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 282.

results are to be generalized.[20] This extension becomes crucially important from the external validity perspective as it justifies generalization.

2.3 What is "natural" or "nature"?

To begin with, in the word-combination "natural experiment", "natural" points to the agent that can be labeled "nature". However, this leaves unexplained what kind of "nature". Without deeper analysis, since "natural experiments" are observational studies, "nature" seems to mean simply that no researcher has explicitly controlled the treatment allocation. The aim of this section is to clarify the features of the force ("nature") that is "in charge". This is needed to simplify the assessment, since the strong and weak points of concrete natural experiments may be a direct consequence of these features.

As I am going to show later, the absence of a human controller does not mean that no human has been involved in the allocation. The involved agents may have affected the treatment allocation significantly, but for them the process was not experimental. First at least initially, the randomness of the treatment assignment was not their aim. Second, there may be factors unrelated to the agents which may have affected the assignment. Other people, for example scientists interested in the case, have considered the conditions of the allocation (e.g. in the historical data) sufficient to claim randomness, or more precisely, the "as-if" treatment condition to be fulfilled.[21]

I claim that three distinct "natures" may take place. Their mixtures are also possible. First, in relation to games, probability theory uses the terms "*a state of nature*" or "*a state of the world*". They refer to a set of external (to the players) conditions at a particular point of time. States of nature occur according to a known distribution.[22] Literally, in its turn "nature" makes a move, giving out a random value, which refers to a set of game conditions, which together constitute "a state of nature". Supposing each "state of nature" equally probable, the "nature" becomes a synonym for perfect randomization with no human control. These are ideal conditions: the "nature" turns out to be an "appropriate" researcher controlling the treatment allocation.

Second, "nature" may be an existing but non-human agent. An example of this is a hurricane, a climate, a flood, etc. The answer to the question whether their behavior can be classified as random is not straightforward. This article is not a paper on

20 William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi- Experimental Designs for Generalized Causal Inference,* 341.

21 This key condition is going to be discussed in the following subsection.

22 Roger B. Myerson, *Game Theory, Analysis of Conflict* (Harvard University Press, 1991), 352.

Climatology or Environmental Science which presents a detailed investigation of such a kind. Still, roughly and intuitively, the behavior of "natural" agents of such natural agents contains both "random" and "non-random" components. What is more important in the analysis is that at least two obstacles can be easily observed disallowing such agents to provide randomization of a treatment. First, the agents are outcomes of the environmental conditions, meaning that a probability, for example, of a flood appearing, is correlated with the characteristics of the area where it happens. In the language of statistics, this means the treatment allocation is not orthogonal to unit features. Second, in terms of inhabitants of the area an endogeneity, or a mutual influence, between effects of the factors and a likelihood of the factors is highly possible. This type of nature does not seem to provide sufficient conditions for a randomization.

Third, a consequence of a complex, often highly detailed, process may be referred to as "nature". As a consequence, the treatment allocation is seemingly randomized, meaning no clear factors influencing the treatment assignment take place. In Political Science, most probably, a multi-stage social or political process may be the "nature". Due to the complexity of a process, both endogeneity and additional covariates are hard to rule out. In this case, "nature" can hardly claim to provide a sufficient randomization.

The typology should probably be developed further. However, even in a very rough form, as presented in this paper, the classification makes the analysis of the validity of natural experiments easier, which will be shown in section 3.

2.4 "As-if" and expected exchangeability treatment condition

In a *natural experiment*, as in an experiment, a treatment assignment must be randomized. Contrary to the latter, in the former the treatment assignment is not random but "as-if" random.[23] This difference appears since a natural experiment is not controlled, but because of conditions it can be claimed that the treatment assignment is "as-if" random. Except for their definition, while dealing with *natural experiments*, the "as-if" part can be easily omitted: natural experiments are so favorable because they are seemingly as randomized as usual experiments.

The condition of treatment randomization, or "as-if" randomization, is the core feature of experimental designs. If the condition is unsatisfied, a design cannot be called an experiment or natural experiment. The idea of natural experiments seems problematic mostly because treatment randomization is extremely difficult to claim in observational studies. Consequently, relaxing the condition of treatment

23 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 283.

randomization and redefining the "as-if" randomization, based on the relaxed requirement, will make the concept of *natural experiments* more justifiable.

In this section, first, I will give a definition of the condition of treatment randomization. Second, I will elaborate on the reasons why the requirements of the definition are crucial for experiments. Third, I will relax the requirements introducing *the expected exchangeability treatment assumption*. Lastly, I will look at the way it modifies the understanding of "as-if" randomization, the condition which enables labeling an observational study a natural experiment.

What is treatment randomization? According to Shadish *et al.*, treatment randomization or random assignment is achieved if "units to conditions are assigned only by chance"[24]. To be more detailed, randomization happens when units of an experiment are divided into control (C) and treatment, or experimental (E), groups randomly, meaning "using some mechanism that assured that each unit was equally likely to be exposed to E as to C"[25] (Rubin 1974, 689). As a result, the averages of all the characteristics of the units are the same in the treatment and control groups. Consequently, the estimation of the treatment effect is unbiased for a given setting. This is the main virtue of randomization of treatment assignment.

The information provided in the previous paragraph can be summarized into the following definition. *Treatment randomization condition is satisfied if, and only if, the treatment is assigned to units unconditional on their features. Each unit has the same probability of being assigned to the treatment.* Now I propose to take a step back and challenge the existing theory by asking two questions: which features of the randomization actually lead to unbiased estimators of an effect? Is the condition of treatment randomization so badly needed? If an effect of interest takes place in a sample consisting of units with expectedly similar features, then which particular units belonging to the control and treatment group should not influence the expected estimator of the effect? In other words, units are interchangeable between the groups, which brings me to the concept of *exchangeability*.

Exchangeability takes place if the joint distribution of a sample is the same as the distribution of every possible permutation of the sample.[26] The effect on a parameter, defined as the difference between the averages of the parameter in treatment and control groups, is a statistic that has a distribution. However, in the case of effect estimation, the main interest is not in the distribution but in the

24 William R. Shadish, Thomas D. Cook, and Donald T. Campbell, *Experimental and Quasi- Experimental Designs for Generalized Causal Inference,* 248.

25 Donald B Rubin, "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology 66(5),* (Oct 1974), 689.

26 John Kingman, "The population structure associated with the Ewens sampling formula," *Theoretical Population Biology 11(2)*, (Apr 1977), 274–283.

expected effect, or in the expectancy of the distribution. Permutations in terms of a sample divided into controls and treatments mean that the units are reassigned.

I propose to introduce *the expected exchangeability condition* that is satisfied if the expected effect of all permutations of the original sample is equal to each other and to that of the original sample. The set of samples satisfying *the expected exchangeability condition* is larger than the set of those satisfying *the treatment randomization condition*, which increases the possibility of justification of natural experiments.

The difference between the conditions can be illustrated with a simple example. Imagine a sample consisting of *n* red and *m* green units. All red units are assigned to the treatment group (*non-randomly*!), but the color is claimed to be irrelevant in the sense of the measured effect. Consequently, despite the violation of the treatment randomization condition, the estimator of the effect is unbiased, since if some green elements are interchanged with the red ones the expected effect does not change, i.e. *the expected exchangeability condition* is satisfied. To sum up, non-random assignment is not a problem if it is conditional only on characteristics, which are orthogonal to the treatment effect.

The expected exchangeability condition is satisfied if, and only if, the reassignment of treatments (within the sample) does not change the expected treatment effect. In the original assignment the units may have different propensities to be assigned to the treatment, but the factors, on which the propensity depends, are orthogonal to the treatment effect. The reassignment may be made after removing or without removing the factors influencing the propensity, after that each unit has the same probability of being assigned to the treatment.

Based on the introduction of the expected exchangeability, a natural experiment can be redefined. *A natural experiment is an observational study, in which the expected exchangeability condition is satisfied.* As seen from the definition, the only advantage natural experiments provide in terms of causal inference (relative to other observational studies) is the "as-if" randomization condition, which can be equaled with the expected exchangeability condition.

2.5 Assessment Algorithm

The practical purpose of this section is to summarize and improve the existing theory to assess natural experiments, or, being more precise, potential natural experiments. Two perspectives are addressed. First, what is the *validity* which is estimated? What "dimensions" does it have? Second, what are these natural experiments? Which of their specific features should be looked at with special attention?

In the following assessment algorithm (AA) these perspectives are reflected along with a critical view of the relevance of the assumed randomization to the compared treatments and controls. I claim that the latter problem is often overlooked in the analysis. The last two points of the algorithm are summarizing.

Steps to evaluate the validity of an assumed natural experiment:

1. Define the type of nature, according to the proposed in (2.2) classification:
    a. As if someone flips a coin ("the ideal case").
    b. Non-human "natural" agent (e.g. a hurricane or a storm)
    c. Multi-stage process (e.g. social or political)
2. Is the "as-if" randomization condition satisfied? How?
    a. Treatment randomization condition
    b. Expected exchangeability condition
3. If the "as-if" condition is satisfied, then answer the question[27] :
    "Is the proposed treatment-control comparison guaranteed to be valid by the as- summed randomization?"
4. Based on the analysis of the previous steps, assess the proposed types of validity:
    a. Internal
    b. External
    c. Construct
    d. Statistical
    e. Content
    f. Ecological
5. Assess the overall success of the natural experiment.
6. Propose ways for improvement (optional).

This plan has both advantages and disadvantages. The main possible pitfall is a high dependence on the applicant. However, the drawback is significantly mitigated by a major benefit, the clear structure. Even if the results of two experts are different, it is easy to compare them. From a certain perspective this disadvantage may become an advantage, since the algorithm provides room for various opinions, keeping them within the set boundaries. Also, to apply the algorithm practically no deep knowledge of statistics is needed, which may be very important for qualitative political scientists.

27 Jasjeet S. Sekhon and Rocio Titiunik, "When Natural Experiments Are Neither Natural Nor Experiments: Lessons from the Use of Redistricting to Estimate the Personal Vote," 2.

3. Empirical Study

3.1 Methodology

The aim of this part is to assess five recent works from Political Science, provided by Dunning (2008, 283), which are claimed to be examples of natural experiments. The cases provide different illustrative examples of possible sources of randomization usual for Political Science: lotteries, differences in laws and regulations, weather conditions, jurisdictional borders, and individual random selection. The major interests of the analysis are to see whether (1) they are "real" natural experiments (2) they can be considered to be able to provide enough information to establish causal links and estimate the effects of interest.

I am going to employ the theory that I developed in the previous section. Relative to the existent literature, three novelties are presented. First, I analyse the plausibility of the "as-if" randomization condition not in the "usual" understanding (randomized treatment assignment), but from the perspective of the newly introduced expected exchangeability condition. Second, I investigate the quality of natural experiments with a focus on possible problems lying "beyond randomization" (points 3 and 4 in the assessment algorithm). I will suppose that the "as-if" condition is perfectly satisfied and look for the caveats which may still be present. Thirdly, I will elaborate on the experiments from the perspective of all types of validity, not only internal and external, but also construct, statistical, ecological, and content.

To prevent possible critiques, a few clarifications need to be made before the presentation of the cases. First, the analysis in this part is very tentative in the sense that the cases are not analyzed deeply to check whether certain types of concrete problems with validity really exist (or not exist) in the case. As stated, the actual validity is a characteristic of an instance of a particular research design application, conditional on certain particular and concretely defined units, treatments, settings, and effects. The aim of the assessments is to point out possible and most probable weak points in natural experiments, or, to be more precise, what are called natural experiments in modern Political Science. Consequently, it is not so important if an actual problem did take place, more important is the high possibility that it may be out there, and so it could appear in a similar case. Second, the purpose of the paper is not to identify all possible caveats in the taken cases but to focus on specific problems. Third, the assessment of the cases does not always strictly follow AA, but uses it as a guideline not as a strict plan. The cases are arranged "in ascending order": each one is better than the previous in terms of validity.

3.2 Case Studies

3.2.1. Effect of affluence on political attitudes

In the first case the randomization is provided by lotteries, which is a rare example of the first type of nature in Political Science. Despite the existence of unfair lotteries, this subsection is not about them. Here the purpose is to investigate the example of the use of supposedly perfectly randomized treatment, fair lotteries.

In his paper Doherty *et al*[28] investigates the impact of personal wealth on individual political attitudes. They employ the lottery wins to rule out the endogeneity between the income and political attitudes. The authors compare the winners of the lotteries with the people from general public via surveys. The major findings of the paper are that the lottery-induced affluence makes the attitudes towards state taxes worse and has less significant effects on the attitudes towards economic stratification and the overall role of the state in the sense of providing social benefits.

In this study the treatment is randomized, so the "as-if" treatment randomization condition can be considered satisfied. As a next step, the question whether the money won in a lottery significantly increases the wealth of a person needs to be asked. Above all, it is not entirely clear which income is meant, relative or absolute. It is probably the relative income that matters. Furthermore, if someone wins money they become richer but not definitely richer than their neighbor. Another weak point is the problem with the construct validity. The authors want to look at the change of wealth on political attitudes, but actually they estimate the effect of the increase of the wealth gained by luck on political attitudes. For example, if someone wins he may start to believe in destiny and less in the welfare state, since he is convinced that he is lucky and does not need state insurance. Importantly, then it is not money that changes his attitudes but the new belief in luck.

More than that, in the case the actual comparison is not between more and less wealthy people, but between the ones who have won and those that have not. To be more precise, the comparison is between the individuals who participated in the lotteries and have won and those who maybe never participated in any lottery and have not won. It is clear that two samples are taken practically from different populations, which completely spoils the internal validity.

The caveats are very significant, since four self-selections take place: the self-

28 Doherty, Daniel, Donald Green, and Alan Gerber, "Personal income and attitudes toward redistribution: A study of lottery winners," *Political Psychology 27(3)*, (Jun 2006), 441–458.

selection to buy the lottery ticket; the self-selection to buy the lottery ticket and win; the self-selection to buy the ticket, win, and participate in the survey; and the self-selection to participate in the survey, while participation in the lotteries is unclear (the control group). The mutual interaction among these self-selections is unclear. For instance, if the participation in the survey is paid then the individuals taking part in it are most likely in need of money and have enough free time, i.e. unemployed, pensioners, or students. If a person wins money and still wants to participate in the survey, then he may be either simply greedy, strongly used to participate, or might be willing to show off. Dunning also mentions the problem of self-selection[29] in his paper.

To conclude, the case has significant problems with all types of validity. Above all, it compares individuals from different populations. In addition, the investigated effect is not one of interest, even if the self-selections are not taken into account. The elaboration on ecological and external validity of any kind makes no sense, since it is impossible to assess the setting in terms of the representation of broader settings, if the setting as such is incorrect. The same can be said about the internal validity: the controls and treatments are simply not comparable. This case may be considered very unsuccessful, and it may not provide any valid causal inference.

3.2.2 Bureaucratic delegation, transparency, and accountability

In his paper Stasavage[30] examines the effect of the transparency of the central bank on the disinflation costs in terms of output and unemployment (2003). The author argues that a higher transparency has a positive economic effect, e.g. the costs of disinflation are lower. In the work the transparency is expressed either in the regular forecasts of the central bank or its reports to the national parliament. According to Stasavage, country-fixed effects are insignificant, and so they are omitted in the analysis.

The source of randomization, "nature", may here be considered of the third type, since the level of the transparency is a consequence of a complicated multi-stage politico-social process. The state institutions and the transparency can be strongly endogenous, which makes the following analysis highly complicated. The treatment is one of the features of a country within the complex inter-temporal systems of effects involving history, current political regime, natural resources, geographical location, and neighboring countries. The transparency can hardly be claimed to be

29 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 285.

30 David Stasavage, "Transparency, democratic accountability, and the economic consequences of monetary institutions," *American Journal of Political Science 47(3)*, (2003), 389–402.

"as-if" random in any sense. Consequently, the case cannot be called a natural experiment.

However, I propose to pretend that the level of the transparency is "as-if" randomly assigned in the setting. Then the answer to the question AA.3 is positive, since the imposed treatment assignment (the level of the transparency of a central bank) is the feature, the effect of which is investigated. However, to rule out other possible treatments in this setting is hardly possible. For instance, some countries may be oil exporters, and then natural shocks of oil price increases are positive for them in terms of output (the taxes rise, social transfers increase, the internal demand increases, and so does output), while for the rest of the countries this shock leads to negative consequences. So, this is not only an example of another factor that may influence output and unemployment, but also the countries are strongly heterogeneous in relation to it.

As was argued in the previous paragraph, the internal validity is doubtful since other factors are not ruled out. The content and statistical validities are good, since the authors use the official economic data. The construct validity is fair: seemingly authors give clear names to the constructs in the design. Without going deeper, it can be said that the ecological and external validity of the case are fair (if the sample is representative). Still, certain problems related to country effects are unavoidable, but such concerns always take place.

To conclude, the major problem of the work is that it is not a natural experiment: the treatment is not random in any sense. Consequently, the validity may be assessed as poor. Dunning does not assess the case in his scale, so the comparison is not possible (2008, 289).[31]

3.2.3 Economic growth and civil conflict

The first case on which I would like to elaborate is the work of Miguel *et al.*[32] In their paper the authors analyse the impact of economic shocks on civil conflicts based on a sample of 41 African countries. According to the authors, the major obstacles in this investigation are presupposed endogeneity and omitted variable bias. Their solution is to use the instrumental variable of rainfall variation in the countries of the sample. The authors argue that, since irrigation systems are not widespread in the area, rain variation has a direct impact on economic growth. Importantly, they do not claim explicitly that their study is a natural experiment. However, the way

31 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 289.

32 Edward Miguel, Shanker Satyanath, and Ernest Sergenti. "Economic shocks and civil conflict: An instrumental variables approach," *Journal of Political Economy 122,* (Aug 2004), 725–753.

they deal with the treatment, rainfall variation, makes it possible to look at the case in this way (Dunning 2008, 284)[33].

Following the assessment algorithm, to start, the "nature" here can be classified as being of the second type. Despite the endogeneity between individuals and weather condition being diminished, since the authors employ not the amount of rainfall but its variation, it may not be eliminated completely. The lack of irrigation is not enough to reject any adaptation to the rainfalls. Even supposing that agriculture is the only way to make a living in the region, the factors influencing wealth cannot be restricted only to rains. For instance, in the areas with a higher rainfall variation the inhabitants may be accustomed to it, and, hence, its impact is less. The possible ways of survival may be reserves or loans in a certain form, etc.

Furthermore, harsher conditions may make people cooperate more, and, in this case, the impact the higher rain variation on social conflicts may be even negative, which is opposite to the main result of the paper. The authors control for ethno-linguistic and religious fractionalization, but do not control for a level of trust in the society. Also, the effect of the rainfall variation may be different for different income levels of population, e.g. heterogeneous, and then overall generalized claims may become highly imprecise in the sense of magnitude. Moreover, heterogeneity is possible on the country level as well. Consequently, the "as-if" randomization condition may be considered unsatisfied: neither as the treatment randomization nor in terms of the expected exchangeability.

In terms of the case, AA.3 could be reformulated as "does the higher rainfall variation give enough grounds to divide the countries into those having bigger or smaller harvests, and as a result, more or less economic development?" Even if the answer is positive, and the exogenous effect exists, it is probably heterogeneous, so generalizations are problematic.

To sum up, I would like to summarize the discussion in terms of validities. The authors do not claim the results of the paper to be generalizable, so it is possible to forget about the ecological validity, and discussing external validity is not needed. The major problems in the case are in internal validity, since the "as-if" randomization condition does not take place. Therefore the overall validity may be assessed as poor. Dunning in his paper does not assess the case.[34] To conclude, the work does not provide any generalization, which makes its contribution limited, and even within the set boundaries, it is very problematic in terms of validities.

3.2.4 Political salience of cultural cleavages

33 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 284.

34 Ibid. 288.

In this case the source of the treatment is the "as-if" arbitrary set jurisdictional border. Posner[35] contrasts the ethnicities of Chewas and Tumbukas in two neighboring countries, Zambia and Malawi. In Zambia the people of the ethnicities are allies, while in Malawi they are adversaries. The major argument of the work is that cultural cleavages between ethnic groups matter in national politics only if the size of the groups is significant relative to the total population of the country. The hostility between the peoples, which results in political tensions as well, was investigated via surveys in a pair of Chewa and Tumbuka villages in the two countries (altogether 4 villages).

The "nature", the source of randomization, in the case is the British colonial authorities. The author argues that the state boundaries are completely random. In reality, the treatment allocation, which is a consequence of a decision-making process, may not have been completely random. However, the factors (if any) which affected the allocation may not be correlated with the investigated effect of ethnic composition. Consequently, the expected exchangeability assumption is satisfied, and the case may be called a natural experiment.

The answer to AA.3 is positive. The treatment and control groups are perfectly comparable in the sense of the treatment. Internal validity may be considered fair, given the strong similarity between the countries, to be more precise, between the pairs of villages. However, the external validity is problematic. Although the results are claimed to be generalizable, the representation of the two pairs of villages relative to the rest of the country and other countries is doubtful. The sample is too small for any generalizations. Furthermore, on the country level other possible reasons for the hostility may not be ruled out completely. Such reasons, despite the overall similarity of countries, may be, for example, different politics of the former dictators or the amount of money borrowed from the IMF.

Additionally, this case has clear problems with methodology. The content and construct validity are questionable and unclear. The author asks certain questions in his interviews and, based on them, makes conclusions about the hostility in the countries. The ecological validity is seemingly weak: the general conclusions are based on the data coming from only two pairs of villages. The representativity of the people according to the population of the countries and to other countries does not look plausible. The statistical validity is good, while the internal one is questionable, since it is impossible to prove the existence of the channel of effect.

35 Daniel N. Posner, "The political salience of cultural difference: Why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi," *American Political Science Review 98(4)*, (Nov 1998), 529–545.

To sum up, this case is good in terms of the "as-if" randomization, but still problematic from many other perspectives. However, at least one improvement is clear: to enlarge the sample of villages. The overall validity of the case can be assessed as fair. Dunning locates the case in the middle of his scale.[36]

3.2.5 The effect of international monitoring on electoral fraud

In her work Hyde[37] looks at the effect of the international observers on the election-day fraud during the presidential elections in Armenia in 2003. The state officials of the country invited them to prove the fairness of the elections on the international level. Since the number of observers was not enough to cover all precincts, they had to allocate their time among randomly chosen voting stations. Their aim was to visit as many precincts as possible. The observers had no idea about the features of the area of the country; consequently, the treatment (their presence) can be supposed to have been assigned randomly. Practically, Hyde estimates the treatment effect on the electoral percentage of the incumbent, or the current president, in a voting station, which is considered to be proportional to the level of fraud.

In this case, the nature may be defined as either of the first type, e.g. random, or the third type, if the observers had certain unknown reasons to pick the precincts, for example, they may have not liked villages the names of which start with "A". However, conditional on the fact that they had no knowledge of the region, then the expected exchangeability assumption may be considered satisfied, meaning that the factors that influenced the choice are uncorrelated with the effect. Consequently, the study is a natural experiment. The next question is whether the decision to visit a precinct actually makes the visit happen. This is true by the definition of the setting of the case.

In terms of internal validity the case can be considered good: the treatment and control units are comparable, the treatment may be considered the only major source of the effect valid for the whole country. However, it is possible to find additional local-specific factors which may have affected the percentage of the incumbent, such as the average local income level. The only possible weak point is the narrowness of the definition of fraud, which diminishes both internal and external validity. For example, theoretically, in different areas of the country the fraud may be in favor of different candidates.

The possibility for a generalization of the effect can be considered positive at least

36 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 285.

37 Susan Hyde, "The Observer Effect in International Politics: Evidence from a Natural Experiment," *World Politics 60(1),* (Oct 2006), 37-63.

for the Post-Soviet region, but only carefully conditional on the country-specific features. The good level of external validity is, above all, the consequence of a high ecological validity. One limitation is that the observers have been invited, and so it is possible that if their presence were imposed the effect would be different. The content validity is good by the precise definition of the settings.

To sum up, the case is a natural experiment: the "as-if" randomization condition is satisfied. The overall validity may be considered good. Unfortunately, Dunning does not locate the study on his scale,[38] and so the comparison is impossible. To improve the case it may be proposed to employ more precinct-specific controls, such as their location and the number of voters.

4. Conclusion

Among five investigated case studies, which have been claimed to be natural experiments, only two satisfied the "as-if" randomization condition. The two sources of "successful" randomness – the random selection by the international observers of which precincts to visit, and the randomly set African state borders – can be summarized as "hidden experiments". In both cases there were individuals who had made random choices, and so actually they were the experimenters. This evidence may lead to the major practical conclusion of the paper that randomization without human intervention is hardly possible: the mysterious nature turns out to be an experimenter who is unaware of his role.

Importantly, in both successful cases, the newly introduced expected exchangeability condition proved to be of great importance. Without "relaxing" the treatment randomization condition, proving the "as-if" randomization may be considered problematic. The expected exchangeability assumption enables ruling out the factors uncorrelated with the causal effect, making the analysis easier.

There are two major sources of problems to the validity. One common threat is the endogeneity in the treatment allocation, especially when the treatment is a consequence of a socio-political process, for instance, state laws. Another problem is the self-selection to treatment. The first case provides an example of severe multiple self-selections, when even a supposedly fruitful source of randomness, the lotteries, does not help to make a study valid.

The paper provides both theoretical and empirical findings. The main theoretical result is the redundancy of the treatment randomization condition: the "as-if" treatment condition may be equaled to the expected exchangeability condition,

38 Thad Dunning, "Improving Causal Inference: Strengths and Limitations of Natural Experiments", 289.

since it is sufficient to provide the unbiasedness of the estimates. The major empirical finding is that many of the assumed natural experiments may not considered natural experiments at all. Interestingly, both "real" natural experiments are "hidden experiments", in other words, the mysterious nature turns out to be human beings.

Bibliography

Banerjee, Abhijit and Lakshmi Iyer. "Colonial Land Tenure, Electoral Competition and Public Goods in India," *Working Paper(2008).*

Brady, Henry E. "Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory," *the Annual Meetings of the Political Methodology Group, University of Washington, Seattle, Washington*, 2002.

Carmines, Edward G. and Richard A. Zeller. "*Reliability and validity assessment*," SAGE University Paper, 1979.

Chow, Yung and Henry Teicher, *Probability Theory. Independence, Interchangeability, Martingales*. Springer Verlag, 1978.

Cox, Gary, Frances Rosenbluth, and Michael F. Thies. "Electoral rules, career ambitions, and party structure: Conservative factions in Japans upper and lower houses," *American Journal of Political Science 44(1)*, (Jan 2000),115-122.

Diamond, Jared and James A. Robinson. *Natural Experiments of History*, Harvard University Press, 2009.

Doherty, Daniel, Donald Green, and Alan Gerber. "Personal income and attitudes toward redistribution: A study of lottery winners," *Political Psychology 27(3)*, (Jun 2006), 441–458.

Dunning, Thad. "No Free Lunch: Natural Experiments and the Construction of Instrumental Variables", *Working Paper (2007).*

_____. "Model Specification in Instrumental Variables Regression," *Political Analysis 16(3)*, (Feb 2008), 290–302.

_____. "Natural and Field Experiments: The Role of Qualitative Methods. Qualitative Methods," *Newsletter of the American Political Science Associations Organized Section on Qualitative Methods 6(2),* 2008.

_____. "Improving Causal Inference: Strengths and Limitations of Natural Experiments," *Political Research Quarterly 61(2)*, (Jun 2008), 282-293.

_____. "Natural Experiments," *Draft entry for the International Encyclopedia of Political Science,* 2010.

_____. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" *in Rethinking Social Inquiry: Diverse Tools, Shared Standards*, ed. David Collier and Henry Brady. Lanham, MD: Rowman and Littlefield, 2010.

_____."Does Blocking Reduce Attrition Bias?" *Newsletter of the Experimental Section of the American Political Science Association,* 2011.

Dunning, Thad and Susan Hyde. "The Analysis of Experimental Data: Comparing

Techniques". *Working Paper* (2008).
Freeman, David. "Statistical models and Shoe Leather," *Sociological Methodology 21*, (1991), 291–313.
Haynes, Stephen N., David C. S. Richard and Edward S. Kubany. "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods", *Psychological Assessment 7(3)*, (Sep 1995), 238-247.
Heathcote, Christopher Robin. *Probability: Elements of the Mathematical Theory*. New York: John Wiley and Son, 2000
Hyde, Susan. "The Observer Effect in International Politics: Evidence from a Natural Experiment," *World Politics 60(1),* (Oct 2006), 37-63.
Kingman, John. "The population structure associated with the Ewens sampling formula," *Theoretical Population Biology 11(2)*, (Apr 1977), 274–283.
Myerson, Roger B. *Game Theory. Analysis of Conflict*. Harvard University Press, 1991.
Miguel, Edward. "Tribe or nation: Nation building and public goods in Kenya versus Tanzania," *World Politics 56(3)*, (Apr 2004), 327–62.
Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. "Economic shocks and civil conflict: An instrumental variables approach," *Journal of Political Economy 122,* (Aug 2004), 725–753.
McDermott, Rose. "Experimental Methodology in Political Science," *Political Analysis 4*, (2002), 325–342.
Posner, Daniel N. "The political salience of cultural difference: Why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi," *American Political Science Review 98(4)*, (Nov 1998), 529–545.
Roe, Brian E. and David R. Just. "Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data," *American Journal of Agricultural Economics 91,* (Dec 2009), 1266-1271.
Rubin, Donald B. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," Journal *of Educational Psychology 66.5* (Oct 1974), 688–701.
Shadish, William R., Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi- Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Company, 2002.
Sekhon, S. Sekhon and Rocio Titiunik. "When Natural Experiments Are Neither Natural Nor Experiments: Lessons from the Use of Redistricting to Estimate the Personal Vote," *Working Paper* (2010).
Stasavage, David. "Transparency, democratic accountability, and the economic consequences of monetary institutions," *American Journal of Political Science 47(3)*, (2003), 389–402.
Trochim, William M. *The Research Methods Knowledge Base, 2nd Edition*. Available at http://www.socialresearchmethods.net/kb/ in September 6, 2011.