NO FREE LUNCH: COSTS AND BENEFITS OF USING THE CONCEPT OF NATURAL EXPERIMENTS IN POLITICAL SCIENCE

By

Elena Labzina

Submitted to

Central European University
Department of Political Sciences

In partial fulfillment of the requirements for the degree of Master of Arts

Supervisor: Tamas Rudas

Budapest, Hungary
(2011)

Abstract

Natural experiment is a research design which is widely employed in modern Political Science. However, in the existing literature the understanding of it is ambiguous. The major aim of the work is to diminish the ambiguity. In the paper the existing theoretical framework on the topic is developed with introducing the notion of *the expected exchangeability treatment assumption* and redefining *the 'as-if' treatment randomization assumption*. First, an overview of the research design methods in the Political Science is made briefly. Second, the concept of a *natural experiment* is critically analyzed and the definition is formulated precisely. Third, the assessment algorithm to evaluate a proposed natural experiment in a sense of its success in terms of the validity is introduced. Lastly, the examples of the applications of the algorithm to seven cases of natural experiments in Political Science are provided.

Contents

Introduction

Recently the concept of *natural experiments* became fashionable and, consequently, widely used in Political Science. Because of the popularity, a growing number of researchers tend to employ it. However, there is still no work where 'natural experiments' at general are analyzed in terms of validity for making causal inference conclusions. What are potential traps and costs of it? What are benefits? The main puzzle of this paper is to see which are the major problems in the natural experiment in terms of validity. Despite various attempts, it is still to a large extent unsolved. The reason of it is, at least, partially a consequence of the ambiguity of the definition of the concept, which needs to be diminished if the problem is to be investigated. Clearly, there is no short sing answer to the puzzle. The paper as a whole is going to be an extended generalization of the existing works on the topic and answer by itself. Meanwhile, certain brief major conclusions are to be made in the end.

Before going further, it is necessary to underline why such an investigation may be needed in Political Science. The investigation will provide a researcher in the field with a sort of a *map* of the minefield of threats of validity. Based on it, before choosing a concrete research design, he or she can estimate the possible problems, based on the provided examples. If the researcher chooses to employ *natural experiments*, then the paper will provide him or her with a lists of potential problems and concerns.

It is going to be shown in the paper that currently *natural experiments*, probably, are over-rated, and the costs of using them very often may outweigh the benefits, which does not mean that natural experiments always provide only threats to validity for sure. The idea is that a researcher

should not always seek for more natural experiments, keeping in mind, that even in the most theoretically (and, personally, hardly achievable) setting, when the *randomization of treatment* was actually provided by the nature, the causal inference may be highly problematic to derive. *Natural experiments* provide challenges of both types, quantitative and qualitative, since "they often take place at the intersection of quantitative and qualitative methods" (Dunning 2008, 283).

The plan of the work is as follows. The paper consists of three parts, each of which represents a step of the investigation of the topic, and the conclusion. The idea is that each section contributes to the answering to the puzzle of the paper by itself, and simultaneously provides a ground for the following sections.

In the first main section I am going to make a general reading overview of the literature on the topics of research designs in Political Science. Then I am going to provide a broader introduction to the topic of constructing the framework for the further analysis. I will give a brief overview on research methods in Political Science; elaborate on types of validity, going beyond the most 'common' *external* and *internal* ones. In the second section I will elaborate particularly on *natural experiments*. In this section, above all, I will give my version of the extended definition, trying to look critically on the concept as such. Then I will discuss the notion of *randomization*, which is crucial in the concept. Why is it usually doubtful to claim that the randomization was provided? Why randomization is so much wanted? Which benefits does it actually brings? I will try to look at the virtues of randomization critically. I will elaborate on the 'as-if' randomization condition and introduce the new concept of the treatment expected exchageability assumption. Then, I will go further and see what happens, if the conditions of randomization are satisfied, which are the most likely sources of threats to validity in natural experiments? In this subsection I will move along the path proposed by Jasjeet S. Sekhon and Rocio Titiunik (2010). In the third part I will employ the results from the previous sections. In this section the aim is to collect the results of the analysis, refining the argument and underlining the most important points. Then I will introduce the assessment algorithm, which can be applied to a proposed case of a natural experiment to evaluate it. As the next step, I will apply the algorithm to seven existing works in the

field, which I take from Dunning's list (2008). Lastly, I will conclude, estimating the approximate leverage of using *natural experiments*, answering the puzzle of the paper.

The main possible caveat of the work is that in many cases, there is no single and correct answer; the choice of a best research is based on specific details of each setting and problem. To diminish the effect of the personal opinion in the paper, I will use the existing works on the topic in each my step. However, especially in the assessment of the cases, the results are significantly dependent on the personal opinion of the author and be scrutinized.

1. Building the framework

1.1 *Reading overview*

A number of works written on the topic of natural experiments in Political Science has been published recently. Jared Diamond from University of California and James Robinson from Harvard University edited, probably, the most significant work on the topic, the book 'Natural Experiments of History' (2009). The following works from the book can be provided as examples of natural experiments.

In 'Shackled to the Past: The Causes and Consequences of Africas Slave Trades"(2008) by Nathan Nunn, the author looks at the relation between the number of exported slaves from an area in Africa and the level of economic development, which he measures in present GDP per capita. The arbitrary, or random variable, is the current country borders, which Nunn claims to be not endogenous with numbers of exported slaves. He found that the areas where the numbers are higher, the current level of development is lower.

Another one is 'Colonial Land Tenure, Electoral Competition and Public Goods in India'(2008) by Abhijit Banerjee and Lakshmi Iyer. Their object of investigation is the connection of original institutions and the current level of public goods. They express the original institutions with the variable of the type of land tenure in the colonial times, landlord-based systems, individual cultivator-based systems, and village-based systems. The authors argue that regions where there used to be the landlord-based system, at present public service providing is worse (meaning less schools and roads). The source of the treatment randomization in the work is the type of the tenure

system, which the authors argue to be arbitrary.

Also, there is the article, where the author is interested exactly in finding natural experiments in Political Science: "Improving Causal Inference Strengths and Limitations of Natural Experiments"(2008) by Thad Dunning. In the text the author lists and analyzes twelve academic works on natural experiments in Political Science and ranks them based on 'plausibility that assignment to treatment is 'as-if' random'. These works provide examples of different causes of 'natural randomization', namely, electoral redistricting (Ansolabehere et al, 2000 and Glazer and Robbins, 1985), precinct consolidation in California gubernatorial recall elections (Brady and McNulty, 2004), cross-sectional and temporal variation in institutional rules in two houses of Japanese parliament (Cox et al, 2000), random assignment of lottery winnings (Doherty 2005 et al), party control of White House in previous elections (Grofman et al, 1998), variation in central banking institutions (Stasavage, 2003), etc.

On the question of research design general methodology, 'Experimental and Quasi-Experimental Designs for Generalized Causal Inference'(2002) by Shadish, Cook and Cambell provides a deep insight, mostly, into various kinds of quasi-experiment designs.

Meanwhile, some works focus more specifically on obstacles in natural experiments and quasi-experiments. For example, 'Does Blocking Reduce Attrition Bias?'(2011) by Dunning provides us with the evidence that matching techniques do not help to get around attrition bias. The author concludes that '[o]ther strategies ,'for example, investing resources in tracking down units that are lost to follow up, or just a random sample of those units (Geng et al. 2008)' may offer better alternatives for reducing attrition bias'. In 'The Analysis of Experimental Data: Comparing Techniques'(2008), he with Susan Hyde compare the relevance of using the parameters of intention-to-treat and the effect of treatment on the the treated in various types of experimental design.

Dunning's another another article, 'No Free Lunch: Natural Experiments and the Construction of Instrumental Variables' (2007), gives an interesting view on one of the most technical aspects of using natural experiments. What he, basically, says is that sometimes if the variables

provided by natural experiments are used, the estimate is not what is really needed. This can happen, if the independent level of interest is endogenous in a model, but there is some exogenous variable (the one provided by a natural experiment) which fulfills two necessary conditions to be used as an IV in the model to measure the impact of the endogenous variable. Then the author agrues that this makes sense, only if the effect is homogeneous, otherwise the estimate is just the impact of the IV variable but has nothing to do with the effect of interest. He develops the argument to a more general level in 'Model Specification in Instrumental-Variables Regression'(2008).

Also, Dunning looks at possible quantitative methods which can be applied to both real experiment and natural experiments to improve their inferential advantages in 'Natural and Field Experiments: The Role of Qualitative Methods'(2008).

The plausibility of alpha and amega of all quantitative research methods, regression analysis, is brought into question in the almost classical 'Statistical models and Shoe Leather'(1991) by David A. Freeman. The authors argues, that despite textbooks usually suppose that '[r]egression usually works, although it is (like anything else) imperfect and may sometimes go wrong, in reality it is something between Regressions sometimes work in the hands of skillful practitioners, but it's suitable for routine use' and 'Regression might work, but it hasn't yet'. In 'Design-Based Inference: Beyond the Pitfalls of Regression Analysis?'(2010) Dunning investigates, whether core problems of regression-based analysis can be overcome with virtues provided by proper research design, and, more specifically, natural experiments.

'When Natural Experiments Are Neither Natural Nor Experiments: Lessons from the Use of Redistricting to Estimate the Personal Vote' by Jasjeet S .Sekhon and Rocio Titiunik(2010) addresses critically the very notion of natural experiments. In their paper the authors not only provide actual examples from current articles, but, which is, probably, more important, try to go 'beyond randomization' in their analysis. They propose to see, which possible problems may still be in place even the treatment randomization condition is satisfied. The main focus of their critical analysis is on the construct validity of the cause-effect channel in the design. Also, they introduce the concept of 'ecological validity', representativity of the setting.

Several authors attempt to look at the causal inference problem from philosophical point of view. For example, Shadish et al are interested in *validity* as such and conditions of the validity. Interestingly, in his paper 'Models of Causal Inference: Going Beyond the Neyman-Rubin-Holland Theory'(2002) Brady Henry investigates various approaches to validity and causal inference.

1.2 *Designs in Political Science*

The construction of the causal inference is one of the core tasks which political scientists need to perform in their research very often. Unfortunately, this goal very often turns out to be a real challenge, since there are numerous sources of bias for the estimates of interest in Social Sciences generally and Political Science particularly. Among the most common causes of the bias are self-selection, non-compliance, dropout (unit attrition), effect heterogeneity, endogeneity, violation of parallel trend assumption, omitted variables and unexpected confoundedness of dependent and independent variables (Shadish et al 2002, 33-102). To rule the sources out the researchers have various tools in hand. However, the question, whether a best research design exists seems debatable: all of the ways of the investigation have its pros and cons.

What is a *research design*? According to KKV , *'[a] research design is a plan that shows, through a discussion of our model and data, how we expect to use our evidence to make inferences' (2004, 118), to put shortly, it is 'the structure of research'* (Trochim 2006).

The dimensions along which methods are usually compared are internal and external validity (Roe and Lust 2009, 1)[1]. Besides that, the important notions here are 'mundane realism' and 'experimental realism' (McDermott 2002, 333). Given the dimensions, I suppose that, a method of a research design can be called 'the best', if it gives a maximization of the sum of internal and external validity over the set of all possible methods, with the extent of its 'realisms' taken into account, for a particular case.

Basically, all methods can be divided into those providing randomization and those that

[1]For more dimensions, meaning validities, see the next subsection

are not. Randomization condition in its most common use, which is employed here, means that treatment was assigned randomly, in other words, that it is not correlated with certain features of the units of investigation or/and other exogenous confound variables. The main reason why randomization is so important is the desire to avoid a selection bias. According to Dunning's definition, random methods are experiments (2008, 282). If the condition of randomization is not satisfied, then methods are divided according to the existence of a control group or multiple measures (Trochim 2006). If the condition is met, then such a method is a quasi-experiment; otherwise, non-experiment.

Theoretically, the best research design is experiment, because of the reasons described in the previous paragraph. Regrettably, in Political Science this design is often unachievable, mostly because (1) the event(s) of interest took place in past, and the researcher has to do with historical data; (2) the scope of the object of the research is so big that it is impossible to run experiments on such a level; (3) there are moral or/and ethical restrictions of performing the experiment; (4) in the case of persons as units the self-selection bias is hardly avoidable. Meanwhile, there is an opinion that experiments are overrated (reference). For instance, one of the most common concerns about experiments, which Rubin underlined a long ago in his paper (1974: 690), is their possible weak external validity.

Initially, experiments could be of two types: *field* and *laboratory*. Then a new type was introduced, *natural experiment*, that is basically just an observational study, in which the researcher can claim the condition of random assignment is satisfied. From its definition it can be noticed, that the newly introduced type seems to be rather artificial, since the belonging to it depends highly on the opinion of the one who classifies. Experiments are so desirable, because theoretically they provide perfect internal validity. At the same time, the level of external validity may be kept, at least theoretically, on the high level as well. At the same time laboratory experiments are criticized because the external validity could be doubtful. Experiment realism of them can be very high, while mundane realism not for sure. Field experiments provide a good balance between both realisms and validities. However, since the researcher intervenes into an investigated process

somehow, the external validity could be violated. Natural experiments potentially provide both very good external and internal validity: randomization and no intervention of a researcher.

'Quasi-experiments' are mostly large-n observational studies with unknown (or not fully known) levels of randomization. The most used methods to achieve randomization with large ns are re-sampling, regression-discontinuity analysis, matching, stratification, and IV-analysis (Shadish et al 2002, 103-243). In the cases of re-sampling and matching, the main problem is that it is difficult to justify that all needed variables have been taken into account: there is always a possibility that some omitted variables have affected a dependent variable of interest. Both internal and external validities can be rather good in these designs: despite the possibility of omitting variable the quality is somehow restricted, given a careful research analysis, a satisfactory level both of representativeness and causal inference can be achieved. The biggest problem of regression-discontinuity is a functional misspecification leading to a decrease of internal validity. Meanwhile, external validity of this method is, probably, the same as of the previous two. The main problem of the IV methods is that perfect instruments do not exist: there is almost always a potential for some correlation which will spoil both internal and external validity of the inference.

Furthermore, it is worth keeping in mind that in terms of a perfect generalized *causal inference*, random allocation of treatment is not the only randomization needed. There are four groups of components in a research design: *treatments*, *outcomes*, *units*, and *settings*. The usual statistically perfect setting consist of two stages of randomization (which are often confused with each other, escpecially in non-statistical literature)(reference). First, a *representative* sample is taken out of *population*. Statistically the representative sample may be obtained if units are taken *randomly* from the population. Second, a treatment is *randomly* allocated over the set of units. How it is seen, here only two groups of components are randomized: units and treatments. The other two, settings and outcomes, are very rarely randomized due to high costs. Randomization of outcomes is technically rather difficult, it involves unachievable bigger amounts of data, and usually researcher randomize other components, taking all outcomes as given. Randomization of settings is closely connected with the construct validity. The question is which causal-effect link is

of interest: the given particular link from a concrete setting or the researcher is trying to establish a more generalized causal relationship.

While performing an experiment or quasi-experiments, the particularity of the setting or settings theoretically should be taken into account. Practically, very often a researcher decides not to bother to think about special features that may have a crucial importance in the causal-effect link of interest. For example, a researcher is interested in the question how extra music classes affect results in the math class in children aged between 10-12. He or she takes data from a few schools and concludes that effect is significantly positive. How can setting be important in this case? The schools, children, country or region of the country, teachers are all the parts of the setting (by the way, there may be much more crucial factors, which should be taken into account too). Let's imagine that the researcher took data from three schools for his investigation. In one school children liked the new music teacher very much, which increased the popularity in students of her husband, who was by a coincidence the math teacher, which increased the popularity of maths: those attending the extra music classes started to more interested in maths, and their result improved. In another school, it can be even simpler: the music teacher was a former math student, he made children more interested in math telling funny stories related to math classes from his past. In the third school, the music teacher started to encourage children to study well to get to a good college, and so those attending the classes started to be more serious about all their classes.

In all the cases, the improvements in the math class were not results of more music classes in terms of music. To be more precise, it was an effect of music teachers. However, their effect hardly can be attributed to the fact that they were music teachers. This example is rather artificial, however, provides the idea, why randomization of settings may be crucially important, and the lack of which may make results of a research virtually senseless. In each school the question to which the researcher obtained (if we suppose that there were no more confound variables which affected the outcome) the answer is the estimation of the effect of these particular teachers.

It is important to touch on the group of non-experimental designs which are actually widely used in Natural, Social and Political Sciences. They employ the purposive sampling strategies. So

the sample, relative to the population is selected non-randomly, and it is possible to say the treatment is allocated non-randomly as well. The purposive sampling strategies differ from 'common' random sampling or random-stratified sampling methods, because a researcher considers important to look at the effect in some specially selected samples. One of the most usual practical examples, when such designs are used often, can be found, for example, in the investigation of the relation between secondhand smoke and lung cancer (Shadish et all 2002, 351).

From another perspective the research methods can be classified into qualitative and quantitative. To simplify, in the qualitative methods the focus of the investigation is on the sense and meaning of certain features of unit of interest, while in quantitative the number of units is larger, and the focus is on computed statistics of the data. For sure, quantitative methods always, except for being studies from a pure statistics perspective, involve a certain amount of qualitative analysis, background of the problem, structured models of the estimated statistics, etc. The perfect examples of a qualitative method are case study and discourse analysis. A "perfect intersection" is context analyst. For example, party manifests are analyzed in terms of words frequencies (a quantitive problem) where words were split into semantic groups by experts (a qualitative problem).

In Political Sciences quantitative and qualitative ideally should be used together to "improve" each other. Quantitative methods are usually related to large-n sample, while qualitative to small-n. However, such an intersection as *natural experiments* requires both analysis. Above all, a lot of qualitative knowledge is needed to conclude that some treatment was actually randomized by nature, and that the outcome of interest is a consequence of the treatments: the researcher needs to rule out all other possible "treatments" and be sure that the treatment preceded the outcome, which can be a difficult task in certain historical and cultural contexts.

1.3 *Types of validity*

Before classifying the types of validity, I would like to elaborate on the very concept of it deeply, since it is one of the ground notions of the causal inference. What is *validity*? Generally

saying, validity is the proximity that a statement is true. Consequently, in terms of the causal inference, it is the proximity that an inference is true. Then in terms of effect estimation, it is the extent to which the estimate of the effect of interest is unbiased in a broad sense, meaning not only the preciseness of the effect's magnitude measurement, but also whether the estimated effect was the one actually investigated. The umbrella definition of all these understanding of validity can be stated as "[a] measuring instrument measures what it is intended to measure" (Carmines and Zeller 1979, 17) and is measured properly.

As it can be seen from the definition, the notion of validity is closely interrelated to the notion of measurement. According to Stevens, '[m]easurement is the assignment of objects or events according to rules' (1951, 22). From this definition it is seen that in the case of Social Science, defining measurement is problematic due to two main reasons. First, in these sciences many 'phenomena to be measured are neither objects nor events'. The phenomena of interest are not the things which are possible to touch (Carmines and Zeller 1979, 9). For instance, social alienation or life satisfaction: these are the examples of such phenomena. Second, as a result of the first reason, the way of measurement may be not straight-forward, often it needs to be redefined or introduced, which may involve additional qualitative analysis, which increases a possibility of expert-related mistakes of design. Natural experiments are an example, when the very calling of a design a natural experiment is a consequence of the way how it is measured and defined. Certain problems in natural experiments can be attributed to features of the common units of the experiment, to persons. They can be influenced not only by the effect itself, but also by understanding that the effect takes place.

Going deeper, the underlying concept of validity is the idea of truth. There are four distinct 'ideas' of truth in the philosophy. However, in terms of validity researchers rarely analyze which version of *truth* they incorporate. The theories of truth are *correspondence theory*, *coherence theory*, *pragmatism*, and *deflationism* (Shadish et al 2002, 35). In correspondence theory a statement is true if it reflects the reality or world. In other worlds, it is an empirical approach, which requires a possibility for a generalization, but only if a statement presupposes it. In terms of natural

experiments, the condition of *reality of a statement* is satisfied (here I do not touch on the problem of construct validity), but the generalization can be problematic. *Coherence theory* says that a statement is true if it is consistent with a certain set of statements or claims. Generally, claims about *natural experiments* may belong to a set of claims to the extent the qualitative analysis of an experiment presupposes it. The problems here are the same as in the case of generalization from the first theory. *Pragmatism* states that that a claim is true if believing in this statement brings more consistency to the existent claims. This theory is not very applicable to *natural experiments*, since, roughly, believing that something has happened only because seemingly something similar has happened is an example of false-science ungrounded claims. Deflationism or the redundancy or minimalist theory of truth postulates that something is said to be true just to make shorter a number of underlying true claims. This theory is purely philosophical and hardly can be applied in the framework of this paper.

Returning to the main topic of this section, importantly, *validity* is not a feature of a design or method: it is a feature of a concrete inference (Shadish et al 2002, 34). That is why assessing methods or method, a natural experiment, in terms of research design is always a certain form of an approximation, which is to be influenced by the opinion of the one who approximates. Despite this, certain generalization can be made. Practically, elaboration on the validity of a design does not result in any strict definitions and assessment, but pointing out possible threats to validity, which follow from distinct features of the design. The fact that only an instance of method application can be precisely assessed in terms of validity, not a type of research design, should be always remembered.

Dimensions of validity The two key dimensions of validity are *internal* and *external* (Roe and Lust 2009, 2). In the following discussion I will presuppose that the aim of a researcher, when he or she chooses a research design for a puzzle of interest, is to maximize both internal and external validity, or to be more precise, a hypothetical sum of the internal and external validity. Consequently, I will assess other types of validity - construct, statistical, ecological, and content according to their contribution to the main two.

To start with the main validities, *internal validity* is about whether the proposed causal relation is true in the instance of the research design, i.e for the sample employed in the experiment in the used setting the controlled treatment leads to the investigated effects. From the explanation, it is seen why it is postulated in the literature that internal validity is the best in the lab experiments: the experimenter controls the treatment allocation and have more possibilities in the design to be sure that it is the proposed treatment that caused the effect. Meanwhile, it is can be concluded that it could be implicitly understood that the channel of the effect is most probably known in the most cases, through it is not necessary. However, if it is not known that it becomes more problematic to rule out all other possible factors except for the treatment.

External validity means the extent to which the proposed relationship is true for the population, to which the effect is proposed to be generalized. In other words, how valid is the claim that from the results obtained in a research design for a sample, which was taken from a population of interest, it follows that the results can be generalized for that population. External validity is a very broad concept, and it is very much dependent on the aim of a research design. Strictly, speaking if there is no need for generalization at all, in other words, the only effect for a given setting, sample, and exactly known treatment is of interest then the external validity is perfect, since the population equals the sample, etc.

Since the notion of generalization is closely connected to notion of external validity, it is worth to say a few words about it here. Targets of generalization can be of the three types (Shadish et al 2002, 83). The most common is *narrow to broad*, when based on the results obtained using the sample taken from a population it is concluded the results take place for the full population. Another one is *broad to narrow* is when, given certain general information about the effect, which is claimed to valid for some population, the question is how valid it is for a subsample of the population. The best example is a medicine, about which it is said that it relieves pain in all people, however, practically, the effect can be (although not for sure) rather specific for a given person. The last type is the generalization *at a similar level*. For example, the results were obtained for men, the question is whether they are still valid for women as well. External validity can be a problem

in experiments, not only because of problems related to a used sample, but because the settings and the treatment of the experiment, especially, in the lab ones may be not very usual.

In literature the common view is the trade-off between between internal and external validities, from which it follows that choosing one a researcher sacrifices another. In the relation to the natural experiment external and internal are both problematic to estimate. However, there is an accordance in the literature that they lay in "the middle of the scale" (e.g Dunning 2008, Roe and Lust 2009). Why is it problematic? On one hand, it is not straight-forward, to which extent a given natural experiment is an experiment, so internal validity is unclear. On the other, as in all observational studies the question of the generality of this particular case should be investigated carefully. Natural experiments (or what is claimed to be natural experiments), especially the ones happened a long ago may have had very special "design conditions" and the generalization of any type becomes problematic.

The definitions of external and internal validities do not have major differences in the literature in the field, however, the definitions of other, auxiliary, types of validity and, even more, possible types as such may differ. While describing statistical and construct validity I will follow the lines of Shaddish et at (2002, 33-103). For the description of *ecological* validity I will employ the ideas of Roe and Lust (2009, 3). To describe the content validity I will use the understanding used by Haynes et al (1995, 239). Above the adopted definitions and understandings I will add certain thoughts related to their contribution to the overall validity and, so that, the success of research experiments and, particularly, the subject of investigation of this paper, natural experiments.

Statistical validity is the most *quantitative* among all validities. It is rather technical in the sense that as such and not supported by any other, it says practically nothing about overall success of the design. Statistical validity is based on the answers to two questions. First, do presumed the effect and cause covary? Second, what is the magnitude of the covariation? It is obvious that a mathematical covariation does not prove any causal inference, however, covariation "in numbers" may be a reason to start a detailed qualitative investigation. In controlled experiments it is a natural consequence of controlled internal validity. For observational studies, and so for natural

experiments as well, statistical validity is a necessary but far not sufficient condition to claim that a research design may be successful. Answering expected wonderings about the reasons why two question are asked (about the existence and magnitude), instead of just one (about the magnitude), I can say that these two questions reflect better two stages on which statistical validity can be important. An answer to the first question may be a very first step: the correlation exists, and then a researcher starts a careful qualitative analysis. If the analysis proves the design does provide grounds to claim the existence of causal-effect relationships, then the researcher may become more interested in the second question. He or she may think about choosing a proper statistical method to 'purify' the effect and to make the estimation of the effect more precise. Statistical validity is closely related to internal validity: it is the technical and mathematical aspect of it. There is no possibility that internal valid design does not provide statistical validity. In social sciences from the perspective of statistical validity the question of measurement can be crucial. It is possible to imagine situation that the way of measurement is chosen to provide an easier calculation of statistical validity.

While statistical validity is closely connected to internal validity, the construct validity is connected to external validity. Construct validity is about the validity of "contructs", the words used to name the parts of a research design. A proper way of defining in terms of its name (and so that meaning) of the effect is crucial for the quality of generalizations and external validity. In political science, when the attitudes are measured, for example, it may very difficult to reflect the needed concepts in proper constructs. Controlling construct validity is theoretically is in a power of the researcher. However, the necessary data may be not feasible in the case of observational studies and natural experiments. Another point from which construct validity may become a weak point of the research is the comparative studies which are based on the existent works in the field. A researcher should be crystal clear that different works do not use the same words for different things. From this point validity is closely related to statistical validity: while calculating covariations a researcher should be sure, that the things underlining the employed numbers are actually those he is interested in. And even if the same things are actually measured, then he or she needs

to check that the units are the same. For example, in one database prices may be given in actual dollars and in the other in deflated for a certain year, in such a case he or she just needs to do reprocess the data.

The notions of the content and ecological validity are rather specific and narrow. However, they enable to focus on certain particular aspects of research design, that may be simply overlooked and lost among others. Ecological validity is the extent to which it is possible to generalize the setting(s) of the design to the universe of all setting of interest. In other words, ecological validity is some sort of representativity of the ”sample” of settings to the ”population” of settings for which the inference is made. This validity is very important in terms of external validity. It is one of the crucial aspects of it. Unfortunately, seemingly this aspect is often overlooked in Political Science. Usually a lot of effort is invested to justify the representativeness of a sample but much less to justify the representative of the settings.

Content validity is the extent to which the way how an element of design is measured is relevant to the constructs of interest. As it is seen, this validity is very much related to the construct validity. While in the construct validity the main problem and interest is to name properly the elements of a design, in content validity it is more technical: the question is how to assess them. If the assessment has already been done, as it most probable is in the case of working with historical data, then the content validity is the extent to which this assessment is precise in terms of tools and measures. This type of validity may be a point of concern in the case of natural experiments, when the researcher has to use already existing data, not always knowing for sure the ways how the components of the data were obtained.

2. On natural experiments

2.1 *Analysis of definitions*

To understand what *natural experiments* are I will first split the name of the concept into two semantic parts, *natural* and *experiments*, and investigate which way each can be defined separately. After that I will look, whether any additional meaning appears, when the parts are taken together, comparing the result with the examples from the literature on the topic.

Experiment What is an *experiment* or, saying more broadly, *experimental approach*? It is one of the key terms and notions of the research design in Political Science. I argue that, it can be considered to have four definitions or understandings: *basic*, *minimal*, *intermediate*, and *maximum*.

1. *The basic definition* is the most intuitive: as Shadish et al write, an act of an experiment is " *[t]o explore the effects of manipulating a variable*" (2002, 507). This understanding does not involve any complicated specification and is very close to the everyday meaning.

2. *The minimal definition*, which is a *randomized experiment* or *randomized controlled experiment*, which has been already mentioned several times in the text, emphasizes the feature that *'the treatment was randomly allocated over the sample of experiment and is controlled'* (Dunning ?, 2). This definition is narrow, formal, and leaves a lot of questions unanswered. What is there also in the experiment except the treatment? Who assigns treatment? What is known about the sample? Should it obtain any features? These questions are crucially important from the perspective of the *internal validity* of an experiment.

3. *The intermediate definition* is more sophisticated consisting of three conditions. First,

the effect of the *treatment* is compared to other treatments, which are called *controls* or *control*, meaning the absence of the treatment. Second, the assignment of controls and treatments are randomized, which is a slightly extended minimal definition. Third, the treatment allocation and other experimental manipulations as well are under a control of the researcher, who performs the experiment (based on Dunning 2008, 282).

It should be underlined explicitly that the intermediate definition presupposes the presence of someone who controls the experiment. I argue that this implicitly means that the *treatment allocation (randomization)* is really random, since it is controlled properly, which enables to rule out, at least expectedly, all possible confounds[1], which are one of the main obstacles to derive a proper causal inference. Despite sounding strange, since *randomization* intuitively is supposed to be random by definition, however, practically it is easy to imagine situations, in which seemingly *random treatment allocation* without any control of the researcher may be confused with *treatment randomization* and is insufficient to establish the cause-effect link. I will return to this problem in more details, when I elaborate deeper on *natural experiments*.

Another important point in the intermediate definition is the emergence of *controls*. While again, intuitively, it may seem that the presence of the treatment presupposes the presence of a control, in reality the absence of the treatment may often mean the presence of other treatments, especially in Political and Social Sciences. Consequently, a simple comparison of those units exposed to the treatment of interest and those that are not, often may not provide sufficient conditions to estimate the treatment effect.

4. *The full or maximum definition* requires one more addition, relative to the intermediate one, which is related to the sample used in the experiment. In the case when a generalized effect is of interest (for example, the effect of watching CNN news on political awareness of US citizens), the sample must be representative of the population over which this generalization is made. Obtaining representative samples in the case of persons is always highly problematic, and the factors, for which it is needed to control are different in each concrete case. Leaving that problem aside,

[1]For definition, for example, see Shadish et al 2002, 506

here it must be said that usually experiments presuppose using randomized samples, in other words samples representative of the population of interest (Shadish 2002, 341). So, the needed addition to the definition is a representative sample incorporated in the experiment. This definition becomes crucially important if the concept of an experiment is looked from the prospective of the extent of the *external validity* of a design.

These definitions gradually add more details to the understanding of the notion of an *experiment*. Probably, the most used in the Social, and Political Science as well, is the intermediate one. I consider crucial to understand in each case what is counted as the experiment and take it into account in the analysis.

Natural I suppose the second part of the term *natural experiments*, *natural*, can be understood as directing to its conjugate word, *nature*. In this case *nature* means the force performing the experiment. The understanding of *nature* as of an *actor* capable of controlling experiment is seemingly ambiguous. I consider three options for the meaning are possible. Surely, mixtures are possible as well. However, in the most cases, as I will show while analyzing the cases from Political Science (Section 3), mostly the nature can be considered belonging to one of the three types.

1. In the part of probability theory related to games, the term *nature* is used in the expression *a state of nature*, which is the same as *a state of the world*, meaning a set of the external conditions at a particular point (of time) of the game which influences the game. The states of nature occur according to a specified distribution function (Myerson 1991, 352). Practically, in its turn the *nature* makes a move, giving out a random value, which is attributed to a number of conditions, which together are referred as the state of the world. To simplify, in this case the nature means just a perfect randomization (with a known distribution), since the way *the nature* transforms through its states ideologically resembles if it based its decisions to transform on flipping a multifaceted asymmetrical coin. Consequently, this can be considered an ideal case when nature does the job of the researcher and does it perfectly (i.e assigns the players' payoffs): the nature can be equalized with the researcher.

2. Another understanding of the *nature* is the one meaning some non-human but actually existing actor (for example, a hurricane or an earthquake). From this perspective the nature becomes related to the environmental meaning of nature. Since weather conditions or natural disasters have a big impact on human beings, they become actual actors. The question whether their behavior can be classified as random is not straight-forward. On one hand, since patterns of the environmental factors are known, they seem to be determinated to a large extent. On the other, according to the definition of a random variable in statistics '*a real valued function defined on a sample space is called a random variable*' (Heathcote ?, 26). In other words, if the way how a variable may change (for example, the amount of precipitations in an area varies according to certain rules) is known then the variable is random. From this perspective the natural impact becomes *random*. Meanwhile, I claim that the problem here is not whether to count the natural impact random or not. The major obstacle here lies in the fact that the impact is not independent, or exogenous (through the natural factor as such may be). In long-term the interactions between human beings and nature may become endogenous, meaning that not only the effect impact of the natural actor affects the persons, but the persons affect the impact as well (for instance, they may change their life conditions based on the occurrence of earthquakes in the area). Consequently, despite a certain amount of uncertainty of future, in this case the behavior of nature can be more or less predicted, while the interactions are more difficult to predict: so, the major concern is that endogeneity[2] is problematic to rule out.

3. The third meaning is the case when the treatment allocation is the consequence of a complex hardly explainable process. As a consequence, the treatment allocation is seemingly randomized, but the distribution function is unknown. In the case of Social Sciences, it is logical to add to possible *natural* processes, social and political ones as well. Due to complexity of the process of the allocation, it is hardly possible to rule out endogeneity here as well. The claim of the random treatment allocation seems hardly probable to prove.

[2]Endogenous variable is a variable which is caused by other variables in the model (Shadish et 2002, 507). Endogeneity is the presence of such variables within the model or setting.

Natural experiments Based on the elaboration on the definitions provided above, it can be concluded that a *natural experiment* is an experiment, in which the role of the researcher is performed by *nature*. Two questions immediately emerge: What is the nature? Which type of experiments do we have? For sure, there is no single answer, and it hugely depends on a concrete setting. However, probably, it is possible to think of certain common sense assumptions which are most probably take place in the cases, which researchers in the field come across.

Intuitively, it is problematic to claim that a full experiment takes place. Probably, in the best case (given the randomization of treatment) it can be claimed that there takes place the basic definition, since not much can be said about the control cases[3].

According to nature, it is claimed that '*the data used in natural experiments come from naturally occurring phenomena actually, in the social sciences, from phenomena that are often the product of social and political forces*' (Dunning 2008, 282). So, it can be concluded that most probably the third type of *nature* takes place, and, consequently, very few can be said about the underlying distribution of the treatment. To put simply, the claim that the treatment is really randomly assigned is rather weak in a general case. This is exactly the problem of 'non-random' randomization mentioned above. However, how I will show while analyzing the cases from Political Science, other types of *nature* are possible to be come across as well.

In relation to the *natural experiments*, the usual definition says that the '*assignment of non-experimental subjects to treatment and control conditions is "as if" random*' (Dunning 2008, 283). This definition looks something between the basic and intermediate definitions, given that "as-if" randomization means randomization, with the difference the role of the researcher is taken by nature (whatever it really is). The definitions of natural experiments, which can be found in the literature emphasize, "*naturally-occurring contrast between a treatment and a comparison condition*" (Shadish et al 2002, 17). The definition is rather vague. However, the natural experiments are used like a usual experiment controlled by nature takes place, while *nature* is understood in its first meaning, which strongly contradicts the conclusions obtained based on putting together distinct

[3]However, there is an least one case when the distribution function is known. This is a case of lotteries (Dunning 2008, 285).

definitions of nature and experiments.

If a researcher looks critically at the notion of a natural experiment, comparing the common sense conclusions about this phenomena and textbook definitions, which are actively used in research, then he or she sees a number of possible question and problems. First, since the understanding of nature is the third one of the proposed definitions, how can it be claimed that *randomization of treatment* really took place? Even if a lot of qualitative research is performed, still this claim seems problematic. Second, there is a clear ambiguity in the term of an experiment used here. Given complex forces are involved, the design, probably, may result in a certain form of endogeneity. The third concern is about the sample. Why were particularly these samples from the population subject to the process? The researcher must understand very precisely for which populations of units, he or she aims to make any conclusions. Otherwise, the derived causal inferences do not make almost any sense.

Generalization of the effect in terms of construct and external validities, even given a well-established internal validity, is a big problem in the case of natural experiments. The questions in the previous paragraph lead us to the discussion of the following section. What is actually this '*as-if*' condition, which is necessary to justify the research design of *natural experiments*?

2.2 *'As-if' treatment randomization*

Before coming *treatment randomization*[4] for the case of *natural experiments*, I am going to look at it from the perspective of 'usual' *experiments*. As it can be seen from the previous section, this condition is crucial in the concept of an *experiment* and the one of the most arguable in *natural experiments* (Dunning 2008). If it is unsatisfied, then the design can not be called an experiment. In this section, first, I will give a definition of randomization. Second, I will elaborate on the reasons why it is needed in the research design. Third, I will introduce a new concept of *expected exchangeability treatment condition*. Lastly, I will look how the emergence of it modifies

[4]It should not be confused with *random sampling*

our understanding of *the condition of* 'As-if' *treatment randomization*.

According to Shadish, *treatment randomization* or *random assignment* is achieved, if '*units to conditions are assigned only by chance*' (2002, 248). To be more precise, the randomization is achieved when the 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*' (Rubin 1974, 689), which can be taken as a definition of the concept. As a result that makes all confounds, both known and unknown for the researcher, to be expectedly the same in the treatment and control groups. Consequently, the estimation of the treatment effect is unbiased for a given setting. That is the main virtue of *randomization of treatment assignment*.

In the literature it is possible to find a number of reasons why *treatment randomization* makes the causal inference easier. First, it makes certain that certain features of the units are not confounded with the treatment or treatment condition. Second, it enables to reduce possible threats to validity, since they are randomly distributed over conditions. Third, the groups are equated in the sense of the characteristics of all variables, known or not, at the pretest level. Fourth, it enables the researcher to control the process in a proper way. Fifth, the error estimates are valid under this condition and uncorrelated with the treatment (Shadish et al 2002, 248). To sum up, all of them are just various grounds for the justification why the estimation of the effect is unbiased.

Now let us take a step back and ask the question: which of the features of the randomization actually results in the unbiased estimated of the effect? Is *the condition of treatment randomization* really needed? Basically, if a researcher is interested in the effect taking place in the sample consisting of objects he considers to be expectedly the same (here I forget about the way the sample is related to the population), then which concrete units of the sample belong to control and treatment groups should not influence the estimate of the effect. In other words, the units should be interchangeable between the groups. That brings us to the statistical concept of *exchangeability*.

Exchangeability In statistical literature the basic version of exchangeability is defined to take place '*as long as the joint distribution of [the sample]* $(X_{\pi 1}, X_{\pi 2}, ..., X_{\pi n})$ *is the same as*

$(X_1, X_2, ..., X_n)$ *for every permutation* πk' (Kingman 1977, 184). The more sophisticated definition implies that n random variables are interchangeable if their joint distribution depends only on $k < n$, where k is the size of the subsample, but not on which particular variables are selected (Chow and Teicher ?, 191). To make it better applicable to treatment measuring procedures in the experimental design, I propose to modify this definition slightly. In terms of probability theory if the sample consists of random variables, then the effect can be defined as a random function of the sample X of a size $n+m$, in which n units are assigned to the the control group and m are assigned to the treatment group, while $\phi^T(.)$ is the function of the treatment effect, and the elements in the groups are numbered tentatively without loss of generality: $f(X) = y(X_T) - y(X_C) = y(x_{t1}, x_{t2}, ..., x_{tn}) - y(x_{c1}, x_{c2}, ..., x_{cm}) = y(\phi^T(x_1, x_2, ..., x_n)) - y(x_{n+1}, x_{n+2}, ..., x_m) = f(\phi^T(X), X)(2.1)$. The expected value of the estimated effect is unbiased if $E(f(\hat{X})) = E(f(X))(2.2)$. Then to make the estimated effect unbiased, it is sufficient to have *the expected exchangeability condition* satisfied, which in case of experiments can be formulated as: $E(f(\hat{X})) = E(f(\hat{X}_{\pi 1})) = E(f(\hat{X}_{\pi 2})) = \cdots = E(f(X))(2.3)$, where $X_{\pi k}$ are all possible permutations of X.

The difference of between this definition and the one of *exchangeability* is that now a researcher is uninterested in the whole distribution, but only in the expectance. This approach increases the possibilities of the assignment into the groups of treatment and control, given certain particular features of the sample, so that the condition of exchangeability is satisfied. The set of samples satisfying the expected exchangeability includes the set of those satisfying the exchangeability condition.

For instance, from the equations above it is seen that (2.3) is satisfied if each unit in the sample has the same distribution function (to be more precise, the same characteristics of all units have the same marginal and the joint distribution as well), then it does not matter, whether the assignment is random or not. However, in such a case it becomes a philosophical question what is the random assignment? A more understandable setting can be formulated as follows. Let us imagine that originally there are n red units and m green units in the sample. Then all red units were assigned the treatment (*non-randomly*!), but the color is claimed to be irrelevant in the sense

of *the expected treatment effect*. So, despite the violation of *the treatment randomization* condition, still the estimated treatment effect is unbiased, while the exchangeability condition is satisfied (the color is irrelevant, so if some green elements are interchanged with red ones nothing changes). Consequently, non-random assignment is not a problem, but only if it is conditional on characteristics which are orthogonal to the treatment effect.

To sum up, the set of samples for which the exchangeability condition[5] is satisfied is larger than that for which the condition of randomization is. Meanwhile, the exchangeability is sufficient to make the estimate of the effect unbiased. Consequently, it is more convenient to check for this condition if we are interested whether a method is valid (e.g the estimate is unbiased).

The discussion of exchangeability condition is very relevant in a relation to *natural experiments*. In their case, the necessary condition is usually referred as *'as If' randomization condition* (Dunning 2008, 283). Given the discussion provided above, such a reference underlines the fact that, actually, it is unknown, whether or not treatment allocation is randomized, but the researcher has reasons to claim the unbiased treatment effect is feasible to estimate, since *the condition of exchangeability* is satisfied.

This question, whether the assignment of treatment was random, is the first question, which the researcher comes across when he or she decides whether to incorporate the concept. Whether called natural experiments are really experiments in terms of treatment randomization is a debatable point, which I will show in the section 3 in detail. Dunning argues that '*one finds marked variation among studies that claim to use natural experiments in the plausibility of this claim*'(?, 5). The exogeneity of the treatment allocation is the first and necessary condition which needs to be satisfied to make this claim realistic (Sekhon and Titiunik 2010, 2). However, in the newly introduced terms a researcher has only to establish that the treatment allocation was uncorrelated with the variables which are correlated with the treatment effect. Such a task is uneasy as well, but seemingly more feasible.

The 'As if' randomization condition is core in the definition of the natural experiments.

[5]Further in the text I will employ *exchangeability* instead of *expected exchangeability*.

One of the possible way of 'rankings' of natural experiments is based on the plausibility of this condition to be satisfied (Dunning 2008, 288). Unfortunately, how it has already been said, the only practical virtue which the treatment randomization, and the newly introduced variation of exchangeability as well, brings is the absence of selection bias in terms of the treatment allocation. Meanwhile, as for natural experiments the necessity of only these conditions can be enough to become an obstacle, since actually *natural experiments* are observational studies. I consider that since in this case the researcher deals with historical data, the plausibility of exogeneity (or partial exogeneity if the condition exchangeability is to be employed) of the treatment is problematic to prove in a general case.

2.3 *Beyond the randomization*

This section is about possible problems in case the condition of '*as-if*' treatment randomization is satisfied. To specify, they are problems in the research design of a *natural experiment*, since that is the only condition which makes an *observational study* a *natural experiment* (for the strict definitions see the next section). These problems seem to be overlooked generally in the methodological articles on natural experiments, with rare exceptions such as the work of Sekhon and Titiunik (2010).

From this perspective *natural experiments* look very similar to usual *experiments*. However, given the fact that the former are observational studies, they have higher *external validity* and lower *internal validity*. Still, usually the settings of *natural experiments* are very specific, and, since cases of natural treatment randomization are uneasy to come across, the external validity may be considered doubtful generally. Consequently, it can be stated that *natural experiments* are problematic in a sense of both external and interval validity. The motivation for this section can be best described with the statement, '*randomization is overrated*', which in the context of the work can be reformulated '*experiments are overrarted*', or, to be more precise, '*natural experiments are overrated*'.

Classification of the problems

The puzzles addressed in Political Science provide numerous possibilities for such problems. Undoubtfully, any classification of them can be considered incomplete and rather artificial. However, I think it is possible to list the major sources of then into five categories. Below I provide such a classification (based on Shadish et al 2002, 249)

1. The first possible group of problems are *temporal*. In the case of *laboratory experiments* for a researcher it is rather easy to be sure that the cause really precedes the effect, since he or she controls the treatment allocation. However, for *natural experiments*, temporal problems may become a real concern. Even if it is possible to claim that the effect is observed later than the cause, still, it is seemingly impossible to point out when the reason of the cause, the real treatment, takes place.

2. The problems from the first group lead to the second one: the possibility for the existence of *the confounds* and *factors unrelated to the cause, but related to the effect*, which are involved in the investigated cause-effect channel. In case of confounding the results of the investigation may stay partially correct, while the latter factors make the results of the original investigation virtually senseless.

3. The third is group is related to the possibility of *maturing* or *regressing* in units, which are not a consequence of exogenous causes, but a feature of the unit as such. For instance, people get old not because of the air pollution. A modification of this is the maturing or regression of the effect when causes and endogenous reasons interact.

4. The fourth group is related to test effects. They may take place in all types of experiments. They can be of two types: test effects, when test or pretest cause certain effects as such, and the violation of SUTVA[6], when interactions between control and treatment groups effect one or both the groups. The example of pretest measurement effects can be a census, which shows the low fertility rates in a town. After that the mayor of the town starts a special economic program to create more incentives to have children. Meanwhile, after the census the citizens find out the

[6]Stable unit treatment value assumption

conclusions of the census, and by themselves become aware of the problem, and start to have more children without even knowing about the mayor's program. If afterwards the mayor decides to measure the effect of the program, the result will be inconclusive: the growth of the fertility rates is not (at least) fully related to the program.

5. The fifth group is related to the possible changes in the instrumentation, when indicated effects are actually related to the different ways of measuring at pretest and posttest. This becomes a bigger problem in the case of natural experiments, where it can be stated as a problem of quality of historical data. In laboratory experiment the researcher can controls the measurement process, which is by definition is problematic, at least for pretest, in natural experiments.

While employing the concept of a *natural experiment* one should not forget that originally, it is an observational study. Consequently, before searching for the problems similar to those provided above, it may meaningful to ask two questions, which Sekhon and Titiunik (2010, 2) formulate in their paper:

1. Is the proposed treatment-control comparison guaranteed to be valid by the assumed randomization?
2. if not, what is the comparison that is guaranteed by the randomization, and how does this comparison relate to the causal effect of interest?

In my further analysis of the cases from Political Science I am going to employ these questions.

2.4 *Extended definition*

In this section I will summarize briefly the discussion of this chapter. I would like to provide here the definition of *a natural experiment*, which I am going to employ in the analysis of the cases in the chapter 3. Also, here I provide the auxiliary definitions, which I need to define the main concept.

Definition Treatment randomization condition is satisfied if and only is the treatment is assigned to units unconditional on their features. Each unit has the same probability of being assigned to the treatment.

Definition Treatment (expected) exchangeability condition is satisfied if and only if the reassignment of treatment to different units (within the sample) does not change the expected treatment effect. In the original assignment the units may have different propensity to be assigned to the treatment, but the factors, on which the propensity may depend, 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.

Definition Treatment "As if" randomization condition is satisfied if and only if exchangeability condition is satisfied.

Definition A *Natural experiment* is an *observational study*, in which *the treatment "as if" randomization condition* is satisfied.

Proposition

The research design of a natural experiment does not provide any additional advantages in terms of the causal inference relative to *observational studies*, except for the satisfied "as-if" randomization condition. Consequently, the overall success of the research needs to be carefully investigated with the condition taken into account. The guidelines for the investigation may be taken from section 5.3 of this chapter.

3. Discussion

3.1 *Preamble*

The aim of this part is to assess a number of the existent works in Political Science and understand whether they can be considered to be providing a sufficient amount of information to establish the cause-effect link and estimate the magnitude of the effect of interest. I am going to investigate seven works from the list of the natural experiments provided by Dunning (2008, 283). In his paper he evaluated the quality of the *natural experiments* found in the modern works in Political Science in terms of the plausibility of the 'as if' randomization assumption, on which I have already elaborated in section 2.2. Dunning mentions that other reasons violating 'the success of natural experiments' exist but does not provide any further discussion or assessment (2008, 290).

In the section I am going to evaluate the quality of the experiments employing the concepts and notions which I describe in the previous theoretical sections of the paper. The three novelties in terms of the analysis (relative to the existent literature on the topic) are to be provided. First, I will look on the plausibility of 'as if' randomization condition, not limited to the 'usual way', e.g equalizing it with the randomized treatment assignment, but from the perspective of the exchangeability assumption. The reason for that is since, as it was shown in subsection 2.2, it is satisfied over a wider sets of possible samples, the plausibility of the satisfaction of the assumption is higher, and, given its sufficiency for the unbiasedness of the effect estimate, it is seemingly rational to require the satisfaction only of it. Second, I will look on the quality of the experiment

going 'beyond randomization', following the lines described in subsection 2.3. It will be proposed to pretend that the 'as if' random condition is perfectly satisfied, and then look for the problems which may still present in the design. This part of the analysis is to a large extent coherent and interrelated with the assessment in terms of validities, but the focus is on the problems unrelated to a 'wrong' or unreliable way of treatment assignment. Thirdly, I will elaborate on the experiments from the perspective of all types of validity, not only the core and most important, internal and external, but also construct, statistical, ecological, and content ones. Lastly, I will evaluate briefly the overall success of the natural experiment.

A few clarifications need to made before going to the examples. First, the analysis in this part is rather tentative in a sense that each case is not going to be analyzed deeply to check whether certain types of concrete problems with validity really exist (or not exist) in the instance. As it was said in section 1.3, the actual validity is a characteristic of an instance of a given research design application, conditional on certain particular and concretely defined units, treatment, settings, and effects. However, the aim of this paper is to point out to possible and most probable caveats in the research design of natural experiments in a rather general case. Consequently, it is not so important, if an actual problem did take place in an investigated case (which is, unfortunately, is impossible to rule out with a 100% probability), more important is that there is a high possibility where it could be there, so it can appear in a similar case. Second, similarly to the previous one, since one of the purposes of the paper is to develop a framework for the analysis of plausibility that a given case is a natural experiment, it is not crucial, if not all possible problems are pointed out for each case. The focus is on possible distinct problems not on the through-out analysis of cases.

In this chapter, first I provide a brief algorithm which I developed to assess the cases of observational studies which are proposed to be natural experiments. Second, I will analyze seven cases takem from the articles of Political Science, applying the algorithm. Third, after the cases I will provide a brief conlusion, summing up the major findings.

3.2 *Assessment algorithm*

1. Define the type of nature, according to the proposed in Section 2.1 classification:
    - 'Nature' as flipping a coin
    - Environmental nature
    - Nature as a natural process;
2. Understand to which extent the 'as-if' randomization assumption is satisfied
    - Treatment randomization assumption
    - Treatment expected exchangeability assumption;
3. If 'as-if' assumption is satisfied, then formulate the question proposed in section 2.3 in terms of the natural experiment:
    - Is the proposed treatment-control comparison guaranteed to be valid by the assumed randomization?
    - Environmental nature
    - Nature as a natural process
4. If the answer is negative, then try to answer the second question from the section:
    - What is the comparison that is guaranteed by the randomization, and how does this comparison relate to the causal effect of interest?
5. Based on the analysis of the previous steps, assess the proposed types of validity:
    - internal
    - external
    - contruct
    - statistical
    - content
    - ecological;
6. Assess the overall success of the natural experiments;
7. Propose ways for improvement (optional).

3.3 *Case studies from Political Science*

3.3.1 Economic growth and civil conflict

The first case which I am going to elaborate on first comes from the paper of Miguel, Satyanath, and Sergenti (2004). The authors look on the impact of economic shocks on civil conflict in 41 African countries. The major obstacles in this investigation are a presupposed endogenity and omitted variables bias. The solution proposed by the authors is to use the instrumental variable of rainfall variation in the countries. In Sub-Saharian Africa irrigation systems are not wide-spread, which, according to the authors, makes the claim of the impact of rain variation on the economic growth credible. As a whole, this research design is not claimed to be a natural experiment by the authors of the article. However, the way they deal with the treatment (rainfall variation) stating that was exogenous gives us a right to think of this setting as of a natural experiment (Dunning 2008, 284).

The part of the research in the paper which I am going to critisize is the influence of the supposedly exogenously given rainfall variation on the economic growth. I am not going to elaborate on the overall claim of the paper about the connection between the economic growth and civil conflict. Meanwhile, If the link between the rainfall variation and economic growth is invalid, then the overall claim becomes invalid as well, which does not mean that there are no other problems to analysis. However, I would like to focus on that part of their analysis, since it is related to the subject of this paper, natural experiments.

To start, I would like to say that it is exactly the rare second type of nature (see section 2.1). In this case the endogenity between the people and weather conditions in long-term is diminished, because the authors employ not the amount of rain, but the variation of the amount. However, the problem of endogeneoty is still possible. Since the rain variation is a feature of a specific region as well. So, saying that irrigation is not wide-spread, authors do not rule out all possible ways of adaptation to the rainfall variation in the region. First, let us suppose that agriculture is

the only way to make a living in the region (later I will relax this assumption), but even it is like that, there are still other possibilities. In the regions with a stronger variation of rains the habits of people should be more appropriate for this, and, so, the impact less. The possible ways to survive in these conditions are habits to have reserves, loans in some form in case of a bad year, etc. Also, if the conditions are harsher (e.g the rainfall has a higher variation) people may have more intensives to cooperate, and then, referring to main question of the paper, the impact of the higher rain variation on social conflict in long-term may be even negative. In the paper authors control for ethnolinguistic and religious fractionalization (Miguel et 2004, 732), but do not control, for instance, for level of trust or cooperation in the society. Second, let us suppose that people have other ways of overcoming problems related to the lack of rain. This preposition seems plausible, since people in the areas of higher variation are supposedly more likely to have alternative options to survive.

It must be said that these propositions about ways to overcome the problems related to rainfall may have impact only on a certain part of population (say, above some income level), however, even so, it makes the overall implied argument of the exogenous nature of the rain variation implausible. Furthermore, the ways of adaptation and the success of it may be probably country-specific, for example, because of different history, for instance. The examples of possible dimensions over which countries may vary may be such as the former country colonizer, the amount of slaves exported from the area in the slave-trade times, or the distance to sea.

Consequently, based on the discussion in the previous two paragraphs, seemingly ”As if” condition is not satisfied. Either in the version of treatment randomization or in terms of exchangeability. So, strictly speaking, even without going beyond randomization assumption, the research does not seem credible in terms of causal inference. However, I will elaborate briefly on the problems besides the treatment allocation.

As I wrote in the subsection 2.3, there are two questions, which need to be asked. The first one is ”is the proposed treatment-control comparison guaranteed to be valid by the assumed randomization?”. In terms of this paper, the question can be reformulated as ”does the rainfall

really make it possible to divide the countries into the ones which has higher harvest in the year or even not harvest but economic growth". Authors try to justify this by saying that irrigation systems are not wide-spread in the areas. However, based on the discussion provided above, it is clear that the comparison is imperfect. The second question is about which comparison is quarantined. That question is not straight-forward to answer. What is known exactly is the variation in rain, but how it affects the particular areas in a sense of economical growth is very problematic to estimate. Even more, it is seemingly implausible, even if the actual (exogeneous) effect exists, to claim that it is homogeneous, how it is implied by the statistical regression method used in the article.

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 generalized, so it is possible to forget about the ecological validity, and discussing exogenous validity is not needed. The major problems of the instance of research design are in internal validity and validities related to it. While the statistical validity is good, which can be seen in the regression results the authors present, the content validity may be problematic. Authors mention problems with the data, which they overcome, but still the way the concepts are employed may be considered questionable. The construct validity is rather good, given the fact that the rain is actually rain, meanwhile, the names of the effects are not clarified precisely.

Dunning in his paper just mentions the work (2008, 288), so, unfortunately it is not possible to compare his judgements with mine. In my opinion, briesfly, the work does not provide any generalization, which makes its contribution to the knowledge limited, but even within the set boundaries, it is very problematic in terms of validities.

3.3.2 Political salience of cultural cleavages

The second case presents an example of one of the major 'sources' of proposed natural experiments: 'as-if' arbitrary set jurisdictional borders. Posner contrasts ethnicity groups, Chewas and Tumbukas, in two neighbor countries Zambia and Malawi (2004). In Zambia the people of the groups are allies, while in Malawi they are adversaries. The major claim of the work is that

cultural cleavages between ethnisity groups matter in terms of politics, only given the significant size of the groups relative to the total population of the country. The hostility between the peoples, which results in the political tensions as well, was investigated via the surveys in a pair of Chewa and Tumbuka villages in the two countries (altogether 4 villages).

The 'nature', meaning the original source of randomization, in the case was the British authorities in the colonial times. It is argued that the set of the boundaries was completely random. It is that, probably, we have the nature of the third type (2.1). The treatment allocation, the allocation to one of the countries, was a result of the process, the logic of which may exist, but now it is not clear. In other words, despite the authors' thought that the allocation was random, meaning not conditional on anything, except for the will of the authorities, in practice there may have been certain factors which influenced the decisions.

Despite the fact that is problematic to conclude, whether the treatment was allocated randomly, based on the discussion provided in the section 2.2, it is possible to claim that that is not crucial, if it is possible to claim that the satisfaction of exchangeability assumption takes place. Remembering the example of the treatment assignment which is non-random and based on the color of unit, then it does not lead to problems with the satisfaction of 'As if' random assumption, since the condition of exchangeability is satisfied. I consider that it is possible to say, taking into the account the justification of 'as if' assumption provided in the paper (surely, in the paper it is just stated that it was random), that the assumption is satisfied. Consequently, now it is needed to look beyond randomization.

The first question from section 2.3, that can be reformulated in this case as "is it true that after the division with the border, the peoples of the ethnicity groups in one of the countries constitute a proportionally small in terms of the overall politically active population share, while in the other country a big one?". The answer to the question is positive, since it is one of the major facts used in the paper. Since the answer to the first question is satisfied, then there is no need for the second question.

Based on the already performed analysis, the case seemingly resemble a real experiment

in terms of treatment allocation. However, the rest of the problems, especially those related to persons and settings, are still in place. I will touch the question of the sample a bit later, when I elaborate on external validity of the case (the results are claimed to be generalizable). Now I want to argue about the possible problems with the setting.

The major problem, as I see, is in the plausibility of the rather narrow and precise claim, that cultural cleavages become politically salient only if the share of the peoples is significant in terms of the overall politically active population. The possible preciseness of the claim seems very problematic, since intuitively it is felt that, probably, there should be other reasons. However, the overall falseness of the claim is not specific for natural experiments, while I would like to focus on such.

The first problem is in the impossibility to rule out other possible reasons of the hostility. This is exactly the problem of uncontrolled settings, which are possible to control in real experiments, but impossible in natural. The possible reasons, despite the overall similarity of countries, can be, for example, the former dictators having different politics or the amount of money borrowed from the international monetary fond. Or, maybe, the reason is more general, and in one country the population is less diverse than in the other, and people in less diverse countries tend to be more sensitive to the ethnicity of persons.

Also, in this design, there are clearly seen problems with the methodology. The content and construct validity are questionable. 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 bad: the general conclusions are based on the data coming from only two pairs of the 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 internal is questionable, since it is impossible to prove the existence of the channel of effect.

To sum up, this case is good in terms of the 'as if' randomization, but still problematic in many other perspectives. However, I see the potential in this: the scope can be investigated further (probably given less restricting claims). Dunning locates the case in the middle of his scale, but

it is important to remember that he equalizes the 'as if' treatment randomization with the 'usual' treatment randomization, which is not totally implausible in this case, but still does not seem, strictly speaking, close to perfect.

3.3.3 Effect of affluence on political attitudes

The third case is about the practically only way of seemingly actually random (in a sense of probability theory) treatment allocation, which is possible to find in Social Sciences. The role of nature here is played by lotteries. Surely, the claim of randomness is valid only if the lottery is supposed to be fair. Probably, there are cases when lotteries are not fair, but in this subsection I am not interested in them. In this subsection the focus is on fair lotteries, and the analysis of which can be applied in some form to any type of fair lottery, not only with money (for example, the order in which skiers start in ski racing, if the order is not dependent on preliminary starts).

In the paper Doherty et al (2006) investigate the impact of lottery winnings on the political attitudes of the individuals. They use the natural experiments of the lotteries to rule out the endogenity, which exists between the income and political attitudes. The problem with investigating their relation is that there are various possibilities for the existence of confounds which influence both attitudes and income. The authors compared the winners of the lotteries with people from the general public via surveys. The major findings of the papers are it follows. Lottery-induced affluence increases negative attitude towards state taxes, significantly increases attitudes towards state redistribution, and has less significant effects on attitudes towards economical stratification and the overall role of the state in providing social benefits. Before starting the analysis, it is must be mentioned that the authors present in the beginning of the paper a list of certain possible caveat of their text (Doherty et al 2006, 443 - 444). However, they still claim that the results are generally valid.

In this case the nature can be considered of the first type (2.1) or actually random in a sense that it is similar if the treatment is allocated with a coin. This provides enough grounds to think that here the 'As if' randomization condition is satisfied as good as it is possible in natural

experiments. It is like a normal experiment, and there is even someone to control the process of drawing numbers. In terms of treatment allocation, it can be even considered an experiment, however, the other parts of the design are still those of a natural experiment.

Going beyond the randomization, let us reformulate the first question in terms of the experiment. Given that practically the money won in the lottery in the study represent a shock to income, ”does money won in the lottery really increase the wealth of a person?”. Without looking into the text of the article, the answer to this question is seemingly unclear. Especially if supposedly, probably, it is a relative change in income not absolute that effects if effects anything. Also, it should be clarified who is to be compared in terms of income. If someone wins money he or see becomes richer relative to his or herself before the winning, but it does not mean that the person becomes richer then his neighbor or anyone else. Another thing which is worth mentioning here is a problem related to construct validity. I consider that the authors want to look on the effect of wealth on political attitude, but here they look at the effect on the *change* of wealth on political attitude or even ”the effect of increase of wealth gained by luck on political attitudes”. These are two (or even three) very different but interrelated cases. The possible example may be that if someone wins money he starts to believe more in density and less in the welfare state, since they believe they are lucky and, so, will not get ill, etc. So here it is possible to say, that the answer to the first question is more ”no”, than ”yes”.

The next question, which comparison is actually provided by the treatment, has the answers: people who won against the ones who do not. Or, to be more precise, people who participated in the lottery and won with the people, who maybe never participated in a single lottery, and do not won. To make it even more precise, it can be said that the people from the control group participated in the lottery with a probability which is equal to the percentage of people buying the lottery tickets. So, even now it is seen that people in the two groups are taken from practically different populations, which spoils hugely both statistical and internal validity. Meanwhile, the problems are even worse. The populations are very diffirent, because there are three self-selections: the self-selection to buy ticket, self-selection to buy ticket(s) and after winning

participate in the survey, and self-selection with a given probability to buy the ticket and after that participate in the survey. The case becomes even worse, because these are three self-selections of various kinds. For example,iIf the survey participation is paid then the people taking part are most-likely either students, old, poor, unemployed, or simply think that it is fun to participate in surveys. If a person wins money and still participates in the survey, then he may be really greedy or be used to participate in such events, or might be willing to show off.

To conclude the analysis of the case, it is seen that the analysis has problems with all types of validity. Above all, that is because it compares individuals from different populations. Also, the problem is that another effect is investigated instead of the one of interest (forgiving about the problem with samples). It makes no sense to elaborate on the ecological and external validity, since it is impossible to assess the setting in terms of representativity of the broader settings (all possible, for instance) if the setting by itself is simply wrong. The same can be said about the internal validity, since the control group is not similar to the treatment group. I consider this case of a research design very unsuccessful and do not see, how it can provide any valid causal inference. Dunning mentions the problem of self-selection (2008,285), but he does not go deeper to see that there are a few different self-selections, which make the analysis practically senseless.

3.3.4 Incentives of Japanese politicians to joint factions

In the fourth case Cox, Rosenbluth, and Thies(2000) investigate the effects of different electoral rules on candidates' and parties' behavior. They want to use the natural experiment of the bicameral structure of Japanese parliament and look at the impact of the electoral rules of each of chambers on factionalism in the chambers (2000, 115). The justification for the effect is that the different rules in the chambers provide different incentives for the parties, which result in the consequences in the terms of 'the party number, size, and the internal structure'(2000, 115).

Interestingly, even the first step of the usual investigation - defining the type of nature - could be considered problematic in the case. Obviously, the treatment is the different rules in the chambers of the parliament, and the investigated effect is the impact of these rules. The problem

is in the question, what is the process of the treatment assignment? Who assigns it? In this setting there is no external 'experimenter' of any type. The problem is that the units, parties or candidates, assign themselves. Consequently, in this case the nature may not defined within the proposed in section 2.1 system. Seemingly, there is no treatment randomization of any kind. The units self-select themselves into the groups. On this stage it may be considered justifiable to stop and state, that it is obviously not a natural experiment, and, so, it is senseless to go further in the analysis. However, I propose to try to reformulate the design to make it is possible to go further.

The only possible units to be taken in order to make it a form of natural experiment are the chambers themselves. If the electoral rules were assigned to them randomly, then we can see the consequences in terms of parties and individuals of the chambers. Given the approach, parties and individuals become themselves part of the effect. Even in such the view the treatment randomization seems implausible. Hardly, it can be supposed that the rules were made up of nowhere. The chambers have different rules and obligations according to the state laws. So, the electoral rules are not the only differences between them. Consequently, it seems that neither the assumption of treatment randomization nor exchangeability assumption may be satisfied. As a result, the necessary 'as if' condition is not satisfied.

Going further, let us pretend that the 'as if' condition is satisfied in terms of the original setting: the parties and individuals are randomly allocated to the chambers. Then the first question (2.3) is whether the allocation into the chambers brings the parties and individual into the conditions of rules of the chambers. This is true by the definition of the design. However, another problem which emerges if we look at the question from this perspective is about the time, when the effect starts to have place. Probably, the conditions of elections should start to effect the units before they get to the parliament. Here we again came to the problem of self-selection and unclarity with the definition of the one who assigns the treatment and then. Seemingly, these problems are avoidable, and they even let pretend that the assignment was random, becoming an obstacle on each stage of the analysis. The possible question to be asked is about the impact of electoral rules on the self-selection process. Another problem is the presence of the SUTVA violation effects: the

same party may be in both chambers of the parliament.

Summing up the natural experiment in terms of validities is difficult. Besides the mentioned problems, the generalization of the case of the Japanese natural experiment is problematic, consequently, the ecological and external validities are unclear. However, the worst part is in the internal validity, because of the self-selection. The construct validity is good, since it is in the definition of the electoral rules. The content validity is fair, because the used figures are actual factual data. The statistical validity is fair as well.

Dunning does not rank the case according to his scale (2008, 289). In my opinion, the thing that can be said surely about the work is, that it is not a natural experiment at all. Most of the problems, both in the evaluation of the research design and the problems of validity in the design as such, come from the misspecification. To have the investigation one probably should look at the case as at a usual observational study taking into account all problems of self-selection and mutual effects between the chambers. The part which is good in the work is the data: the electoral rules and the fact about parties and individual are factual and precise and disallow any mistake in reading.

3.3.5 The effects of international monitoring on electoral fraud

In the fifth case the effect of the presence of international observers on election-day fraud is under the investigation. In her work Hyde looks at the presidential elections in Armenia in 2003 (2005). The country state officials invite international representatives to prove the fairness of the election. Since the number of the observers is not enough to be in all precincts during all the time of the election, they have to allocate their time between randomly chosen polling stations and spend in each a certain short amount of time to be able to visit as many precincts as it is possible. The observes have no idea about the area and features of precincts, consequently, the treatment (the international observers) is claimed to be assigned randomly. It is supposed that with the observers present in the polling station the fraud is impossible. Hyde's aim is to estimate the treatment effect on the electoral percentage of the incumbent, or the current president, in the precinct.

In this case the nature may be considered either of the first type, if we imagine that the observers flip a coin to choose the precincts, or the third time, if actually they have certain reasons to choose the polling stations, for example, they do not like the ones starting with "W" (2.1). However, given the fact that the international representatives have no information about the region to be true, then it is exactly the same as in the example of the color-depend treatment assignment (2.2). Consequently, the condition of exchangeability may be considered satisfied, and, so, the 'as if' condition satisfied as well. So, this case can be considered a natural experiment.

The first question is about whether the decision to visit a certain precinct actually chooses the visit to happen. This is true by the definition of the natural experiment. The problem is that it is unclear, which effect is measured, the effect of their presence in the exact time they are there (so the fraud is impossible only and exactly in that time) or the deterrence effect, which lasts longer, maybe the whole day. If it is the deterrence effect then there is a possibility that the fact of the presence of the international observers as such in the country has effect not only on the precincts which the observes visit. If it is known in advance, that there is a positive probability that they may visit, it may have an effect by itself (suppose that the people in the polling stations which are visited did not about that in advance).

In terms of internal validity the experiment can be considered fairy good: there is seemingly no reason to suppose that there are other common, not precinct-specific reasons, which may have effected the fraud. However, what is actually measured is not the fraud itself, but the percentage of the support of incumbent in terms of voters. This obviously can be influenced by various reasons. Maybe, in some regions the incumbent is supported more that in the other. For example, if the incumbent is a communist, and the village generally support the communistic ideas, then they may support him only because of belonging to the party. The problem of equalizing the votes for the incumbent with the fraud diminishes the construct validity, which means threats to both internal and external validities.

The possibility to generalize the effect can be considered questionable, but possible to some extent, for instance, to the Post-Soviet region, but only conditional on the local-specific features.

This is possible because of a fair level of ecological validity. However, a limitation is that the observers are invited, so, probably, if the international presence was imposed then the effect would be different. The content validity is good, since it is presupposed by the setting: all the data is exact. The statistical validity is good as well, the estimates are significant. However, the value of their significance is diminished by the fact it is not exactly explained, which effects are measured.

To sum up, the case is a true natural experiment, since the 'as if' randomization condition is satisfied; the overall validity of it may be considered not bad. Unfortunately, Dunning does not locate the case on his scale (2008, 289), so it is impossible to compare. The proposed way to improve the experiment may be taking into account the precinct-specific characteristics. For instance, their location, the size in terms of number of possible voters, and the type of the institution they are, such as a school building, university, etc.

3.3.6 Bureaucratic delegation, transparency, and accountability

The sixth case of my analysis can be classified both as belonging to Political Science and Economics as well. I will follow Dunning and will look at it as at an investigation of Political Science (2008, 283). In his paper Stasavage examines the effect of the transparency of the central bank on the disinflation costs in terms of output and unemployment (2003). The major claim is that the higher transparency, expressed either in regular forecasts or reports to national parliaments or both, has a positive effect, meaning that the costs of disinflation are lower. The source of the natural experiment in this case is the differences in the transparency of the central banks between the countries. According to the author, country fixed effects are insignificant, and based on this estimate, he states that the only difference between the countries of the analysis is the level of transparency of their central banks.

Starting from the source of randomization for the case, it can be said that the 'nature' here is seemingly close to the third type (2.1), since the level of the transparency of a central bank is obviously a consequence of a certain politico-social process in the country. The process was probably multi-level and extremely complicated in its structure. Also, the strong institutional en-

dogeneity is supposedly present in the setting. The level of transparency of the central bank is one of the features of the country within the complicated inter-temporal system of effects involving history, current political regime, natural resources, neighbors, etc. It is seemingly highly implausible to claim the level transparency is assigned exogenously. Consequently, the case may not be considered to be 'a natural experiment'.

However, I propose to pretend that the level of transparency is 'as if' randomly assigned in the setting. Then the answer to the first question 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 countries this shock leads to negative consequences. So, that 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 it was said in the previous paragraph, the internal validity is problematic, since other factors are not ruled out. The content and statistical validities are good, since the authors use the economical apparatus which is very precise. The construct validity is fair: seemingly authors give clear names to the constructs in the design. Without going deep, it can be said that the ecological and external validity of the case may be fair, if the sample is representative of the population of countries (or, alternatively, is the population itself). However, the problems related to country effects are avoidable, but such a concern is usual, so this is not a drawback of the research.

To sum up, the major problem of the work is that it is not a natural experiment, despite it is claimed to be. If it is not, then the whole setting and approach becomes highly problematic in terms of the causal inference. Dunning does not assess the case in his scale, so the comparison between his and my views does not seem possible (2008, 289). The proposed improvement may be a certain form of controlling for the state institutional system as a whole or to employ carefully chosen very similar countries with the only difference in the level of transparency of the central

bank.

3.3.7 Nation building and public goods provision

The seventh, and the last, case is about the impact of nation-building state policies on the diminishing of the negative effect of the ethnical diversity on local public goods outcomes. Miguel in his paper investigates the case of Kenya and Tanzania (2004). While in Tanzania the state pursue various such policies, '*most notably in language use, education, and local institutional design*'(2004, 1), Kenya does many less policies of the kind. The author says that, since the countries are very similar in terms of historical and cultural backgrounds and geographical charactristics, they can be considered similar in terms of everything except for the treatment, which is in this case the nation-building policies.

In terms of randomization this case is very similar to the one presented in 4.1.2. The countries of the analysis are both former colonies. However, after a closer look, there is a huge difference. Here, given that the division of the countries was 'as if' random, the units of comparison, the countries in their present state may differ not only in the terms of nation-building state policies. These policies emerged as a consequence of a politico-social process, similar to the one from the previous case. Probably, they are one of the numerous factors, which effect the life in the country. Consequently, in this case there may be significant problems, first, in terms of the 'as-if' random assignment of lands to the countries. Second, even if we believe that these policies are randomly assignment, it is hard to believe that the countries are the same, except for the policies. So, this case can hardly be classified as a natural experiment.

At the next step, if we pretend that the 'as-if' randomization condition is satisfied, it is needed to formulate the first question from section 2.3 in terms of the case. In the setting there is no problems with the question since the assignment of the policies means just the assignment of policies. It can be considered that the construct valitidy is fairy good. Because of the severe problems with randomization, the internal validity may be considered problematic. The channel of the effect is unclear, the treatment may be not exogenous.

The ecological and external validity as a whole are moderate: the sample of countries is very small, since only two are compared. Even if they are very representative, which can not be said surely, the results from such a small sample can hardly be actually generalized to the whole region of Africa. The content is difficult to assess, however, the authors use precise numbers of statistics of local goods outcomes, so it can be considered fair.

To sum, this is as problematic case as the previous one is: it has significant obstacles to be called a natural experiment. The ways of improvement are similar to the ones proposed for the previous case: to try to control for more differences between the countries. Dunning mentions this case in his paper (2008, 286), but does not locate it on his scale (2008, 289).

3.4 *Summing up the cases*

To sum up the analysis of the cases, it can be said that the results are overall expected in the sense of the major problems found. Usually it is considered in the literature, that the most problematic and crucial in the concept of natural experiments is the 'as-if' randomization assumption, which appeared to be the biggest problem in five out of seven investigated cases, even in the relaxed form of the exchangeability assumption. Consequently, according to the definition, which I provided in section 2.3, only these two cases can be called actual natural experiments.

I consider it is worth mentioning the sources of the randomization in these cases. In the first one, it is the jurisdictional borders, borders of two countries in Africa, which emerged as a result of the will of the representatives of the British Empire. In the second one, it is the choice made by the international observers which polling statitions to visit during the presidential elections in Armenia in 2003, given the fact that they are unaware of the specific features of the regions of the country. In both cases the introduced exchangeability assumption helped to justify the 'as-if' assumption, even if the randomization assumption may be possibly violated.

Despite the fact, that the rest of the cases are not natural experiments, it is still useful to look at other problems in the cases (besides the violation of 'as-if' assumption and the problem of

the endogeneity of the treatment). Since the object of the investigation of the paper is more not the actual natural experiments and even not the observational studies called natural experiments, but the way how the concept used and understood in the contemporary Political Science. Briefly, the major problems can be devided into three categories.

The first category is the problems related to the samples used in the natural experiments. In the first paper, authors knowing that their sample is not representative simply avoid to claim about the possibilities to generalize. In the second one, the sample is not representative according to the population of the proposed generalization. In the third control and treatment groups come from different populations, and there are severe problems related to self-selection. Overall, it is seen that problems with self-selection are one of the major sources to the threats to validity in the investigated 'natural experiments'. The second category is the problems related to the settings, which are very problematic to generalize for a broader case, or the ecological validity of the experiments. The cases took place in such a specific settings with many unknown factors involved, that even if the condition of 'as-if' randomization is satisfied, generalization is problematic to justify. The last but not least category is problems in the way how the components of the design are defined and measured, to be more precise, the problems related to construct, content, and statistical validities. Overall it can be said that the natural experiments which were analyzed are neither good in terms of internal, nor external validities. However, it does not mean that all the cases should be considered to be completely hopeless in terms of the causal inference. For instance, the second example can be significantly improved.

It is seen that many of the problems usual to natural experiments are a consequence of that they are simply observational studies. However, the impression is that a researcher, when he or she has found a 'natural experiment', e.g a treatment which is seemingly 'as-if' randomly assigned, he or she forgets about other typical problems, such as confounds and effect heterogeneity. Even if a treatment is 'as-if' randomly assigned, that does not rule out other possible problems. Meanwhile, from the perspective of the 'as-if' randomization assumption, the hugest problem is self-selection of the units and overall endogeneity in the unit-effect-treatment relation. To put briefly, it is needed

to clarify before any further analysis, whether the units were involved in the treatment allocation themselves.

To conclude, the proposed tools in terms of analyzing the cases of natural experiments are proven to be useful and powerful. Following the same steps, which are easy to remember, the analysis can performed quickly and precisely. Also the advent of such an analytical algorithm makes comparison of the cases very feasible and clear.

Conclusion

The major contributions of the paper can be summorized as follows.

First, I investigate the existing publications related to the natural experiment and find the incoherence and impreciseness in the theory about natural experiments. Second, summing up the existing knowledge and explicitly developing the theory, especially, for the crucial 'as-if' randomization assumption, introducing the new concept of the expected exchangeability assumption, I make the theory in the field more precise and coherent.

Based on the contribution in the sense of theory, I propose a clear and concrete algorithm for the assessment of the instances of the research design. Despite the fact, that the algorithm was introduced to be applied to natural experiments, with slight modifications it can be applied to any design. Surely, it may be applied not only to the existent already investigated cases, but also in the beginning of the analysis to rule out possible problems and locate the proposed research in terms of validity.

Last, I provide the examples of the application of the algorithm to the existent natural experiments. Given the examples, I summarize the major problems, which I found in the summary of the section.

References

Banerjee,Abhijit and Lakshmi Iyer. 2008. Colonial Land Tenure, Electoral Competition and Public Goods in India. Working paper.

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

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

Chow, Yung and Henry Teicher. 1978.*Probability Theory. Independence, Interchangeability, Martingales*. Berlin-Heidelberg-New York, Springer-Verlag

Cox, Gary, Frances Rosenbluth, and Michael F. Thies. 2000. Electoral rules, career ambitions, and party structure: Conservative factions in Japans upper and lower houses.*American Journal of Political Science* 44:115-22.

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

Doherty, Daniel, Donald Green, and Alan Gerber. 2006. Personal income and attitudes toward redistribution: A study of lottery winners. *Political Psychology* 27 (3): 441-58. (Earlier version circulated as a working paper, Institution for Social and Policy Studies, Yale University, New Haven, CT, 2005)

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

Dunning, Thad. 2008. Model Specification in Instrumental-Variables Regression. *Political Analysis* 16 (3): 290-302.

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

Dunning, Thad and Susan Hyde. 2008. The Analysis of Experimental Data: Comparing Tech-

niques. Working paper. Presented at the annual meetings of the American Political Science Association, Boston, MA, August 29-September 1, 2008.

Dunning, Thad. 2008. Improving Causal Inference : Strengths and Limitations of Natural Experiments. *Political Research Quarterly 2008 61: 282*

Dunning, Thad. 2011. Does Blocking Reduce Attrition Bias? *Newsletter of the Experimental Section of the American Political Science Association*. Dunning, Thad. 2010. Design-Based Inference: Beyond the Pitfalls of Regression Analysis? In David Collier and Henry Brady, eds., *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman & Littlefield, 2nd edition.

Dunning, Thad. ?. Natural Experiments. Draft entry for the *International Encyclopedia of Political Science*

Freeman, David A. 1991. Statistical models and Shoe Leather. *Sociological Methodology*, Volume 21(1991), 291-313

Heathcote, C.R. 2000. *Probability: Elements of the Mathematical Theory* John Wiley Son, New York.

Hyde, Susan.2006.Can international election observers deter international fraud? Evidence from a natural experiment. Chap. 7 of *Observing norms: Causes and consequences of internationally monitored elections*. Ph.D. diss., Department of Political Science, University of California, San Diego.

Kingman, J. F. C. 1977. The population structure associated with the Ewens sampling formula. *Theoretical Population Biology*: 274-283

Roger B. Myerson.1991. *Game Theory Analysis of Conflict*, Harvard University Press

Miguel, Edward. 2004. Tribe or nation: Nation building and public goods in Kenya versus Tanzania. *World Politics* 56 (3): 327-62.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. 'Economic shocks and civil conflict: An instrumental variables approach'. *Journal of Political Economy* 122:725-53.

McDermott, Rose.2002.Experimental Methodology in Political Science. *Political Analysis*. 10

(4): 325-342

Nunn, Nathan. 2008. Shackled to the Past: The Causes and Consequences of Africas Slave Trade'. Chapter 5. *Natural Experiments of History* edited by Jared Diamond and James A. Robinson (or available as a Working paper).

Posner, Daniel N. 2004. The political salience of cultural difference: Why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi'. *American Political Science Review* 98 (4): 529-45.

Roe, Brian E. and David R. Just. 2009. Internal and External Validity in Economics Research: Tradeoffs between Experiments, Field Experiments, Natural Experiments and Field Data. Accepted to the 2009 Proceedings Issue of *American Journal of Agricultural Economics*

Rubin, Donald B. 1974. Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology. 1974, Vol. 66., No. 5, 688-701*

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

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston - New York. Houghton Mifflin Company.

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

Trochim, William M. 2006. The Research Methods Knowledge Base, 2nd Edition. http://www.socialresearchmethods.net/kb/destypes.php