NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Ilya Lokshin

WHATEVER EXPLAINS WHATEVER: THE DUHEM-QUINE THESIS AND CONVENTIONAL QUANTITATIVE METHODS IN POLITICAL SCIENCE

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: POLITICAL SCIENCE
WP BRP 23/PS/2015

Ilya Lokshin[1]

WHATEVER EXPLAINS WHATEVER: THE DUHEM-QUINE THESIS AND CONVENTIONAL QUANTITATIVE METHODS IN POLITICAL SCIENCE[2]

The paper (for the first time, to the best knowledge of the author) applies the Duhem-Quine thesis to conventional quantitative methods in political science. As a result, the discussion of methodological problems associated with these methods is implanted into the epistemological issues highlighted by the Duhem-Quine thesis. Such a link between popular political science methods and philosophy of science could help clarify the difficulties of the former and give impetus to the improved research practices.

JEL Classification: Y90

Keywords: the Duhem-Quine thesis, quantitative methods, methodology, epistemology.

[1] National Research University Higher School of Economics. Political Science Department; Laboratory for Qualitative and Quantitative Analysis of Political Regimes, Researcher. E-mail: ilokshin@hse.ru.

[2] This Working Paper is an output of a research project implemented by Laboratory for Qualitative and Quantitative Analysis of Political Regimes within NRU HSE's Annual Thematic Plan for basic and applied research. Project title: «The Impact of Feedback Loops in State-Society System on the Resource Redistribution and Economic Growth: Perspectives in Democratic, Autocratic and Hybrid Regimes». The author is grateful to Mikhail Ilyin (Higher School of Economics) and Irina Soboleva (Higher School of Economics, Columbia University in the City of New York) for their valuable comments.

I. Introduction

The progress of modern political science largely depends on finding better quality data and employing more reliable methods of data analysis. Although political science has witnessed great methodological progress over the last 50-60 years, many methodological, epistemological and ontological issues persist[3]. Some of these problems can be well formulated and analyzed using achievements from the philosophy of science. In this paper, I attempt to highlight some of the problems of conventional quantitative methods using the Duhem-Quine thesis, which is an important contribution to the philosophy of science in the 20th century. I do not argue that the Duhem-Quine thesis can shed *new* light on the problems of conventional and the most popular quantitative methods in political science, but I instead claim that 1) it is a comfortable point of departure for analyzing and clarifying these issues; 2) that the thesis can deepen our understanding of these problems by implanting it in the philosophy of science, 3) and that it draws our attention to the seriousness of these problems by revealing their problematic epistemological basis.

There are three caveats to this.

First, I do *not* argue that all quantitative methods are flawed and I doubt that the Duhem-Quine thesis can be applied to all of them. I focus primarily on the simplest and most widely used statistical and econometric procedures, such as significance tests and the not particularly sophisticated versions of regression analysis. However, the basic character of these methods might imply that a large proportion of quantitative methods have inherited the "sins" of the most simple procedures.

Second, the problems which I will highlight are not new to political scientists. My aim is to demonstrate that these problems have deep roots in epistemological issues which are clarified by the Duhem-Quine thesis, and, because these issues are serious, they cannot be easily ignored. Moreover, I do not know of a single book or a paper which would apply the Duhem-Quine thesis and its implications to methodological practices in political science. It may be that I have researched this insufficiently thoroughly, but if I am correct, the absence (or, at any rate, the lack) of attention to the Duhem-Quine thesis in political science should be amended, not only due to its prominent role in the modern philosophy of science, but also because of its usefulness

[3] There is a genuine *Methodenstreit* in political science: see, e.g., King G., Keohane R., Verba S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press; Brady H., Collier D. (Eds.). (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield Publishers; Gerring J. (2012). *Social Science Methodology: A Unified Framework.* Cambridge, Cambridge University Press. For the challenges in the ontological sphere, see Hay C. (2006). Political Ontology. In Robert E. Goodin and Charles Tilly (Eds.) *The Oxford Handbook of Contextual Political Analysis* (pp. 78-96). Oxford: Oxford University Press.

for clarifying the nature of challenges we come across in the methodological and epistemological sphere.

Third, although I focus only on quantitative methods, I do not imply that the qualitative approach is free of issues. However, the more "scientific" reputation that quantitative methods have makes it more pressing to demonstrate that this reputation to a great extent is not watertight.

This paper is structured as follows: the second section describes the Duhem-Quine thesis and some of its implications, demonstrating how relevant the thesis is for the difficulties of political science methodology. The third section discusses some practices of conventional quantitative methods and has two aims; a) to identify some flaws on a purely methodological level (which are generally well-known), b) to demonstrate their link on a deeper epistemological level using the Duhem-Quine thesis's terminology. The fourth section concludes the main ideas of the paper.

II. The Duhem-Quine thesis

II.1. Statements of the thesis

The thesis which later became known as the Duhem-Quine thesis was originally only associated with Pierre Duhem. In a book entitled "La Théorie Physique. Son Objet et sa Structure" published in 1906, Duhem argued: "...[T]he physicist can never subject an isolated hypothesis to the experimental test, but only a whole group of hypotheses; when the experiment contradicts the predictions, what the physicist learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed"[4]. Therefore, Duhem's thesis states that we cannot test and reject an isolated hypothesis but only the whole group of hypotheses and assumptions; the silence of an experiment upon which hypothesis should be modified became known as the "holist underdetermination" of a theory[5].

In general, it is acknowledged that the concept of holist underdetermination is not confined to physics and may arise in various contexts where hypotheses are tested.

[4] Duhem P. (1906). *La théorie physique. Son objet et sa structure*. Paris: Chevalier & Riviére, p. 307.
[5] Stanford K. (2013). *Underdetermination of Scientific Theory*. Retrieved August 18 2014 from Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/scientific-underdetermination/.

In 1954, Willard Quine, before Duhem's book was published in English, came up with a very similar idea to Duhem's thesis, although it had another intellectual foundation and employed different terms and concepts. In "Two Dogmas of Empiricism", Quine argued that[6];

> ...[T]otal science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field. Truth values have to be redistributed over some of our statements <...> But the total field is so undermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to revaluate in the light of any single contrary experience.

The idea that a single fact can be explained in various ways is not particularly new; for instance, a very similar idea was expressed by J.S. Mill[7]. The novelty of Quine's thesis lies in the strong argument that there is no method associated with objects' ontological properties , or with the truth, or with any other absolute substances, which could distinguish between better and worse explanations. The choice of a theory is entirely pragmatic and is guided by the logic of convenience and explanatory success[8].

Clearly, Duhem's and Quine's theses are intimately related to each other. The notion of different ways to redistribute truth values in Quine's analysis is analogical to the statement that we can choose different ways of modifying the group of hypotheses which are rendered false by an experiment.

However, Quine's part of the thesis is primarily associated with the issue of empirically equivalent hypotheses and the absence of purely rational grounds of preferring one theory to another. This difficulty is known as "contrastive underdetermination"[9].

Another implication of Quine's thesis that is worth noting is the blurring of the boundaries between synthetic and analytic statements, meaning that facts are theory-laden. Quine's thesis casts serious doubts over the whole positivistic project of science. There is no clear border between facts and their interpretations.

[6] Quine W. (1951). Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review*, *60*(1), p. 39-40.
[7] Mill J. (1974). *A System of Logic Ratiocinative and Inductive*, p. 500. Toronto: The University of Toronto Press.
[8] Quine W. Op. cit, p. 41-43.
[9] Stanford K. Op. cit.

II.2. The Duhem-Quine thesis, structural underdetermination and political science

For the following analysis, I will introduce another kind of underdetermination, which can be referred to as "structural underdetermination". It occurs when two or more factors may explain a phenomenon ("experience", in Quine's terms) but a) it is not necessary that they contradict one another (as assumed in the case of contrastive underdetermination) and b) a scientist cannot distribute the "explanatory weights" between these factors on purely rational grounds. This is the source of underdetermination, and is called "structural" because it concerns the phenomenon's causal structure[10]. The condition (a) makes it clear that contrastive underdetermination can be perceived as a special (extreme) case of structural underdetermination, with different patterns of truth values as different factors which can explain the phenomenon.

With a considerable effort and rich imagination, political scientists are able to formulate innumerable theories through which whatever can explain whatever.

This is possibly due to 1) epistemological reasons and 2) the character of the objects that political science deals with.

Epistemological reasons can be highlighted using the Duhem-Quine thesis:

a) When evidence proves to contradict a theory, the latter can almost always be modified in such a way that its correspondence with facts is restored. This may be done by the revision of causal links implied by the theory, or introducing some additional conditions to the theory, which specify when the theory holds, or in another way. In any case, a scientist can "redistribute truth values" in the theory and make it congruent with the facts. The problem is that this can be done in multiple ways, and revisions of this kind may be perceived as rather *ad hoc*.

b) As long as there is no clear-cut border between facts and its interpretations, according to Quine's thesis, a single fact can have multiple interpretations which, in turn, can have little in common. As a result, different theories can arise from the same sets of facts. The consequences

[10] What I mean by "the causal structure of a phenomenon" has an intuitive character to a great extent. It consists of "explanatory weights" associated with factors which (weights), if fully known, could precisely predict everything that the scholar wants to know about the phenomenon. The problem is that they are never fully known in political science. However, the notion of "explanatory weight" has to be clarified as well. Intuitively, the explanatory weight of factor A in the causal structure of phenomenon B depends on the three following elements: 1) the number N_A of causal paths leading to B in which A participates (let such a path be named p_A); 2) the relative frequency F_{p_A} when p_A is involved in the causal process; 3) the effect size E_{i_A} associated with the causal path p_A. Then the explanatory weight of A in the causal structure of B can be defined as follows: $\sum_{p_A=1}^{N_A} F_{p_A} \times E_{p_A}$.

of this problem and some examples of it in political science will be examined further in the paper.

Another source of "blooming, buzzing confusion" in the world of theories in political science is the difficulty of the object itself. In society, almost everything is linked to everything else. This simple fact paves the way for the possibility of interpreting statistical connections in causal terms. There are many theories about the causes of democratization[11], democratic peace[12] or resource curse[13], which are typical situations rather than being exceptions. The problem is that all these theories may be true to some extent, but we do not know to *what* extent. Therefore, the problem of structural underdetermination becomes exacerbated.

III. Conventional practices in quantitative methods and the Duhem-Quine thesis

Some methodological practices which are widely employed in a quantitative paradigm intensify the problems of the structural underdetermination of political science theories. In the following sections I focus on three practices of conventional quantitative analysis: 1) significance tests, 2) the underestimation of causal complexity and the causal heterogeneity of political phenomena and 3) dealing with phenomena with a very broad and general character.

III.1. The abuse of significance tests

In economics and psychology, the methodological issue of abusing significance tests is well-known. In economics, McCloskey is the main opponent of significance tests[14] and, although some of her statements have been called into question, others are beyond reasonable doubt. In psychology, the tradition of criticizing significance tests is long-lasting and robust; Jacob Cohen and Paul Meehl's contributions are perhaps particularly powerful and important[15]. However, in political science the issue of significance tests tends to be ignored and cases of criticism are very rare[16]. Nevertheless, the issue is worth discussing. In the following sections I point out two

[11] Мельвиль А.Ю., Стукал Д.К. (2011). Условия демократии и пределы демократизации. *Полис (Политические исследования), 3*, с. 165-169.
[12] Rosato S. (2003). The Flawed Logic of Democratic Peace Theory. *American Political Science Review, 97*(4), p. 585-602.
[13] Ross M. (2001). Does Oil Hinder Democracy? *World Politics, 53*, p. 325-361.
[14] See, for example: McCloskey D. (1985). The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests. *American Economic Review, 75*(2), p. 201-205.
[15] Cohen J. (1994). The Earth Is Round (p < 0.05). *American Psychologist, 49*(12), p. 997-1003; Cohen J. (1990). Things I have Learned (So Far). *American Psychologist, 45*(12), p. 1304-1312; Meehl P. (1967). Testing Theories in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*(2), p. 103-115; Meehl P. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defence and Two Principles What Warrant It. *Psychological Inquiry, 1*(2), p. 108-141.
[16] I have found only 4 papers which criticize the common practices of dealing with significance tests in an explicit way: Gill J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly, 52*(3), p. 647-654; Rainey C. *Testing Hypotheses of No Meaningful Effect.* Retrieved August 22 2014, from http://www.polmeth.wustl.edu/media/Paper/nme.pdf; Ward M., Greenhill B., Bakke K. (2010). The Perils of Policy by p-value.

difficulties with, or misunderstandings of, significance tests and show how the whole issue is linked with the epistemological challenges of The Duhem-Quine thesis.

A. There is (almost) no reason to test null hypotheses in social sciences because a nil effect size (almost) never occurs.

Significance tests are primarily conducted by testing null hypotheses, but null hypotheses imply that the link between parameters is literally zero. Clearly, this assumption is quite unrealistic.

Whether the null hypothesis is rejected or not depends on four factors; the critical level of significance, effect size, statistical power and sample size[17]. If effect size, or the correlation between the parameters, as its special caseis hardly ever null, then statistically significant results can be achieved with a sufficiently large sample. Statistically, testing null hypotheses in most cases cannot be used as a straightforward way of analyzing effect sizes.

The problem is exacerbated by the epistemological issue, Even if the theory is not underpinned by statistical evidence, a researcher, as the Duhem-Quine thesis indicates, can propose a number of alternative theories by modifying some parts of it. These modifications can arise from the internal structure of a theory or the technique of analysis. For instance, every time a theory fails to pass the challenge of the null hypothesis testing, it could be concluded that this is due to a small sample size[18]. Therefore, the peculiarity of null hypotheses significance testing offers additional ways to elaborate on empirically equivalent theories. The capability of scholars to reject theories and be able to choose between them diminishes.

B. Significance tests reveal information about statistical significance but do not tell us anything directly about political or substantial significance.

Statistical significance is often perceived as simply "significance" which is simply wrong. For instance, the coefficient in a regression model can be very close to 0 and practically unimportant yet still be statistically significant. Statistical significance is neither a necessary nor sufficient condition of substantial significance. However, in a large proportion of political science studies, there is no distinction drawn between the two notions of significance.

Journal of Peace Research, *47*(4), p. 363-375; Schrodt P. (2014). Seven Deadly Sins of Contemporary Quantitative Political Analysis. *Journal of Peace Research*, *51*(2), p. 287-300. However, the real problem is not the small quantity of the articles on the theme but the weakness of the resonance they engender.

[17] Cohen J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), p. 156.

[18] Zumbo B., Hubley A. (1998). A Note on Misconceptions Concerning Prospective and Retrospective Power. *The Statistician*, *47*(2), p. 387.

Some critics such as McCloskey even claim that statistical significance is completely useless due to this difference and should be expunged from scientific enterprise. However, this opinion may be contested. Statistical significance shows which estimates are most reliable by differentiating between those which contain too much noise and those in which noise can be largely ignored[19]. When an estimate is not statistically significant, we usually expect a substantial change in it, in a more representative sample. Therefore, statistical significance is what necessary in order to be able to draw reliable conclusions.

Despite this, a further question arises, as to whether statistical significance is the only criterion which allows us to judge the reliability of an estimate. I do not attempt to give a full answer to this question, but there are cases when researchers may be fairly certain that even a small sample is representative, provided that they know fairly well how observations are distributed in the population. Given this, a small sample and the absence of statistical significance do not imply that the estimate is unreliable. Moreover, it is in the domain of political science where such a situationarises quite often.

Many samples used in political science can be seen as populations, not samples, such as a set of independent nations, a set of post-communist countries, the states of the USA, the OECD countries or the parties participating in elections. Nevertheless, it could be counter-argued that all these examples are not entirely legitimate, and could be described as samples drawn from a kind of a "super-population"[20]. For example, the population of independent nations as of 2014 is a sample from a super-population of independent nations from 1950-2050.

Despite this, the concept of the super-population does not make up for the difficulty mentioned earlier. Political scientists are often able to estimate the distribution of observations or cases even in the super-population and therefore are able to judge the representativeness of their sample. Even if the sample is quite small, it can still be reliably representative.

Aside from this, significance tests are associated with critical levels which are widely perceived as "objective" thresholds while in fact they are only useful conventions. The idea behind these critical levels in social sciences should not be automatically borrowed from

[19] Hoover K., Siegler M. (2008). Sound and Fury: McCloskey and Significance Testing in Economics. *Journal of Economic Methodology*, *15*(1), p. 15-16.
[20] Western B., Jackman S. (1994). Bayesian Inference for Comparative Research. *The American Political Science Review*, *88*(2), p. 413-414; Berk R., Western B., Weiss R. (1995). Statistical Inference for Apparent Populations. *Sociological Methodology*, *25*, p. 421-423.

engineering or hard sciences. Social scientists do not decide whether the estimate is significant or insignificant, but instead correct the degree of certitude that a hypothesis is true[21].

If these considerations - about the fact that 1) significance tests are only one of the sources of judgment about the reliability of estimates and 2) about the misunderstandings associated with critical levels - are generally correct, the role of significance tests is further diminished, as well as the simple observation that they do not indicate substantial significance.

In regression analysis, the blending of statistical and substantial significance can easily lead to the underestimation of factors which have statistically insignificant coefficients and the overestimation of those with statistically significant coefficients. As a result, our conception of the causal structure of the phenomenon under consideration can easily go awry.

Another consequence of the abuses of significance tests may lie in a purely theoretical field and influence the way in which scientists formulate research questions. Significance tests imply the oversimplifications of formulations of problems because they allow for only two variations; the estimate is either significant or not. Perhaps the way in which problems are often stated in political science has its roots in this dichotomy. Indeed, for many years researchers have been attempting to reveal whether democracy leads to better economic performance; whether "free resources", such as oil or gas or diamonds prevent democracy and stimulate civic wars; which factors pave the way for revolutions or whether the parliamentary system is better than the presidential, and so forth. The very abstractness of these questions is an issue. If theories behind these questions are correct, under certain conditions and to some extent the answer to these questionsis "yes".Within the mainstream methodological paradigm of social sciences, the questions should not start with "do's" or "whether's", and they should not be formulated in the most general sense ("does democracy affect economic growth?") but with the words "to what extent" does factor X cause phenomenon A and under which circumstances is the effect greater or smaller.

Therefore, a number of practices associated with significance tests can easily lead to the distortion of the causal structure of the phenomenon under consideration and subsequently, aggravates the problem of structural underdetermination. Moreover, these practices sometimes draw attention to not very useful questions and distract from the important issues.

III.2. The underestimation of the causal heterogeneity of the phenomena examined

[21] Rozeboom W. (1960). The Fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin*, *57*(5), p. 420.

Political science aspires to a great extent to establish universal lawlake regularities The problem however, is that these regularities often are rather similar to the Holy Roman Empire: as just as the Holy Roman Empire, as Voltaire argued, was neither Roman, nor holy, nor an empire, universal regularities in political science may be neither universal, nor regularities.

First, it is useful to differentiate between the two senses of the notion "universal". On the one hand, this may mean "applicable to all cases (of the predefined set)". On the other hand, in political science practices it often actually means "applicable to an average case"[22]. The difference between these two senses of "universality" has far-reaching consequences. Universality in the second case may only poorly describe most real cases or observations.

For example, there is a universal regularity which states that democracy does indeed cause economic growth, but what does this actually mean? Does this imply that democracy is equally good for economic growth in all cases, particularly if the sample includes all the independent countries in the world? Or do we assume that its effect for Switzerland would be different for Congo? Further, do we expect that Congo Kinshasa and Congo Brazzaville are identical in terms of the causal link or not? Can we be sure that the causality is valid for all contexts and for all places and for all times? This result cannot be automatically generalized with sufficient reliability to all observations in the sample, or even to most of them. The more causally heterogeneous the set of observations is, the more scant our knowledge is. Therefore, in many cases it may be reasonable to sacrifice universality for the sake of the possibility of drawing more information from the data.

The very fact that the results of an analysis of universal regularities can be applied better to some observations but not to others may again lead to the distortion of the causal structure of the phenomenon we are examining. Moreover, causal heterogeneity automatically means an increase in the number of alternative theories which pretend to explain the phenomenon, thus exacerbating the problem of structural underdetermination even further. The difficulty of the causal structure for very heterogeneous samples exceeds our capabilities to clarify and refine it.

III.3. An analysis of broad and general phenomena

It is fairly easy to test theories on the same level of abstractness upon which they are formulated. For instance, if it is postulated that there is a link between democracy and oil resources, then variables called "democracy" and "oil resources" would be defined and used in the research design.

[22] Long ago psychologist David Bakan made a similar distiction between "general" and "aggregate": Bakan D. (1966). The Test of Significance in Psychological Research. *Psychological Bulletin*, *66*(6), p. 433.

However, the connection between the theory and testing it, which is provided by this strategy, has a reverse side the more general and broad the phenomenon is, the easier it is to develop multiple theories which are empirically equivalent. The reason is that broad and general categories, *ceteris paribus*, can be theoretically linked to more factors than more concrete and narrow concepts. As a result, if the empirical testing of the theory contains very broad and general concepts and is not rejected, we cannot unpack the explanatory weight of different factors which may be associated with the theory. The causal structure of a phenomenon is therefore underdetermined.

The same problem arises when the operationalization of a variable can be interpreted in multiple ways. The result of operatonalization is usually perceived as piece of data, a sort of *fact* upon which theory testing is based. However, as the Duhem-Quine thesis states *in abstracto*, facts and their interpretations do not have clear-cut boundaries between them, which is exactly what we can observe in political science *in concreto*.

By way of illustration, consider two theories linking modernization with democratization. One was developed by Daron Acemoglu and James Robinson[23], the other was developed by Ronald Inglehart and Christian Welzel[24].

Acemoglu and Robinson propose a causal link that can be schematically and in a simplified way rendered as follows: modernization and economic growth -> the decrease in income inequality and the change of economic structure, the shares of physical and human capital increase, average incomes and living standards rise -> the loss that elites can suffer from economic redistribution diminishes -> elites become less disposed to struggle against democratization -> the conditions for democratization mature[25].

The connection between modernization and democratization is analyzed by Inglehart and Welzel as well. For them, the causal path runs as follows: modernization and economic growth -> across society new resources are spreading which a) give impetus to the rise of the subjective valuation of freedom, b) increase the capabilities of citizens to struggle against the elites -> democratic values are spreading -> citizens' pressure the elites for democratization and, due to the new resources, is quite effective -> democratization occurs[26].

[23] Acemoglu D., Robinson J. (2006). *Economic Origins of Dictatorship and Democracy*. Cambridge: Cambridge University Press.
[24] Welzel C., Inglehart R. (2008). The Role of Ordinary People in Democratization. *Journal of Democracy*, *19*(1), p. 126-140; Welzel C. (2009). Theories of Democratization. In Ronald F. Inglehart, Christian Haerpfer, Patrick Bernhagen, Christian Welzel (Eds.) *Democratization* (pp. 74-90). Oxford: Oxford University Press.
[25] Acemoglu D., Robinson J. Op. cit., p. 285-286; 319-320.
[26] Welzel C. Op. cit.

Once again, the fundamental observation behind both theories is the link between modernization and democratization. But the "middle terms" in the causal links are very different. The juxtaposition of these theories reveals at least two problems.

First, from the initial observation there is no way of differentiating between these two theories. In this sense, they are empirically equivalent. It may be possible to distinguish between them with new data, but even then, and if one of these theories should prove to be inconsistent with the data, a theory can be revised so that its correspondence with the evidence could be restored without a radical change of its main propositions. As Quine would say, "truth values" can be redistributed. Therefore, the theories' empirical equivalence may persist even when new data is unearthed. The more ways of modifying theories without changing their core propositions (the more "flexible" theories are), the more serious the danger is of empirical equivalence. At the same time, the more general concepts a theory uses, the greater amount of possible causal paths it can propose to link its two extreme points, the more flexible a theory is.

Second, it is evident that these theories are not particularly compatible.. Acemoglu and Robinson stress that elites play a crucial role in the democratization process, whereas Inglehart and Welzel emphasize the significance of masses. Perhaps the inconsistency is not purely logical, but it is not clear whether the underdetermination we encounter is structural or contrastive, where the latter is worse and more problematic, by implying that one of the alternative theories is wrong.

This example reveals the source of underdetermination associated with the variation of multiple causal links, which are easy to draw when theories discuss very general and broad categories. Another example clarifies the perils of the operationalization of broad categories.

Consider the theory by Carles Boix, which aims to clarify the mechanics of democratization as well[27]. The general assumption is that democracy leads to massive redistribution which is not in the interests of elites. One of the theory's causal mechanisms states that economic growth increases the share of mobile capital in the economy. The elites with mobile capital may move their assets abroad and avoid the (negative) consequences of redistribution, but this option is not open to the elites with immobile capital. As a result, they must oppose democracy more strongly than if they had mobile capital. The more prominent role mobile capital plays in the structure of economy, the better the prospects for democracy are[28].

[27] Boix C. (2003). *Democracy and Redistirbution*. Cambridge: Cambridge University Press.
[28] Ibid., p. 12-13.

What we are interested in here is not the theory as such, but the way key variables are operationalized in the process of empirical (quantitative) testing. Carles Boix uses the following ways to operationalize the variable of mobile capital: the GDP share of the agricultural sector; the share of fuel resources in exports and average years of schooling (it is expected that education creates human capital, which is more mobile than physical capital)[29].

All three ways of operationalization give broad and very different interpretations. The share of the agricultural sector in the economy can be seen as the extent to which the economy is innovative and modern, or even as the degree of society's closeness to the traditional order. The share of fuel exports is perfectly compatible with the interpretation from the resource curse theory; it is the measure of "free money" which may lead to the deadlock and degeneration of democracy. Average years of schooling are usually interpreted in fairly straightforwardly as the degree to which a population is educated. This view of the variable is in accordance with Inglehart and Lipset's hypotheses, which explain the same link between modernization and democratization, albeit in a very different way.

Richness of imagination can provide other interpretations of the operatiolizations at hand. The previous examples are sufficient to show that the ways of operationalizing variables once again lead to the problem of empirically equivalent theories. Under these conditions, the positive results of hypotheses testing cannot increase our certitude that one of these theories is the main driver behind the empirical pattern, not to mention the explanatory weights that we can ascribe to different causal mechanisms. Evidently, there is a problem of structural underdetermination.

It is clearly easier to point out this problem rather than to solve it. A tentative conclusion may be that it might be useful, at least occasionally, 1) to find new data which could help differentiate between theories; 2) to test hypotheses employing concrete and not essentially contested concepts[30], making it harder to link it to a large body of very different causal factors; 3) to stick to the most straightforward interpretations of operationalizations and to restrict scientific imagination; 4) to develop new and more suitable operationalizations instead of adding a new and "forced" sense to old ones and 5) to decrease the heterogeneity of the sample, so as to use narrower theoretical concepts that are more context appropriate.

IV. Conclusion

[29] See for the similar discussion of variable operationalizations in Boix's book: Geddes B. What Causes Democratization? (2007). In Carles Boix and Susan C. Stokes (Eds.) *The Oxford Handbook of Comparative Politics* (p. 323-324). Oxford: Oxford University Press.
[30] Gallie W. (1956). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, *56*(1), p. 167-198.

The aim of this paper was to show how some epistemological difficulties associated with the Duhem-Quine thesis are reproduced and aggravated in conventional practices of quantitative analysis. I focused primarily on 1) the implications of using and abusing significance tests, 2) the temptation to conduct large-N research with causally heterogeneous observations and 3) the direct link between theories and empirical tests, where these tests used operationalizations of broad categories exposed to multiple and very different interpretations. These practices intensify the issue of the structural underdetermination of theories. They do not help clarify the explanatory weights of factors behind the phenomenon examined, and also muddle them and distort the representation of the causal structure.

The problem is aggravated further. As Quine argued, once he had formulated his thesis, the choice of competing theories must be guided by purely pragmatic criteria. However, it is not at all clear which criteria are pragmatic in the conventional quantitative methods of political science. Trade-offs are ubiquitous, the simplicity of regression models is often unrealistic and may easily lead to unfounded conclusions[31]. The paradise of quasi-experiment design, promised by control variables, is not what it seems and a strategy based on control variables is theoretically unsound[32]. Moreover, the temptation to receive straightforward and well-defined results by significance tests suppresses the deeply problematic character of this kind of testing (the challenge was partly discussed above).

The consequences of these and many other trade-offs (or even misunderstandings) is that *many* models pretend to be adequately compatible with the data and researchers do not have particularly reliable methods to establish which models are pragmatically better. In this case, to what extent are these methods scientific?

Given all these difficulties of conventional quantitative methods, and many others which are not covered in this paper, the question remains as to what can and should be done. It goes without saying that the answer to this question lies beyond the scope of this study. John Gerring pointed out the difficulties associated with clarifying causal mechanisms and proposed testing them "to the extent that it is feasible"[33]. It is not entirely clear whether this principle would demotivate scholars to develop new strategies of empirical testing or whether it would decrease the standards of scientific rigor. However, I believe that political science is not in a desperate

[31] Achen C. (2005). Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong. *Conflict Management and Peace Science*, 22, p. 327-339; Schrodt P. Op. cit.
[32] Clarke K. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, 22, p. 341-352.
[33] Gerring J. (2010). Causal Mechanisms: Yes, But... *Comparative Political Studies*, *43*(11), p. 1518. Many ideas on causal mechanisms from the present article have direct links with what Gerring argues on the same theme, however, Gerring didn't rely in his discussion on the Duhem-Quine thesis.

state, and there are many ways which could propose at least a tentative and partial remedy, not only for the challenges of the Duhem-Quine thesis but also for many other issues. Experiments[34], Bayesian statistics, even refined and more sophisticated techniques built upon conventional regression analysis and other approaches can offer a number of alternatives. There is no doubt that it is easier to be "comfortably numb" in a methodological sense, but it is contradictory to the spirit of scientific enterprise.

[34] In the cited article John Gerring champions experiments as well.

References

Acemoglu D., Robinson J. (2006). *Economic Origins of Dictatorship and Democracy*. Cambridge: Cambridge University Press.

Achen C. (2005). Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong. *Conflict Management and Peace Science*, 22, p. 327-339.

Bakan D. (1966). The Test of Significance in Psychological Research. *Psychological Bulletin*, *66*(6), p. 423-437.

Berk R., Western B., Weiss R. (1995). Statistical Inference for Apparent Populations. *Sociological Methodology*, *25*, p. 421-458.

Boix C. (2003). *Democracy and Redistirbution*. Cambridge: Cambridge University Press.

Brady H., Collier D. (Eds.). (2004). *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman and Littlefield Publishers.

Clarke K. (2005). The Phantom Menace: Omitted Variable Bias in Econometric Research. *Conflict Management and Peace Science*, 22, p. 341-352.

Cohen J. (1990). Things I have Learned (So Far). *American Psychologist, 45*(12), p. 1304-1312;

Cohen J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), p. 155-159.

Cohen J. (1994). The Earth Is Round ($p < 0.05$). *American Psychologist*, *49*(12), p. 997-1003;

Duhem P. (1906). *La théorie physique. Son objet et sa structure*. Paris: Chevalier & Riviére.

Gallie W. (1956). Essentially Contested Concepts. *Proceedings of the Aristotelian Society*, *56*(1), p. 167-198.

Geddes B. What Causes Democratization? (2007). In Carles Boix and Susan C. Stokes (Eds.) *The Oxford Handbook of Comparative Politics* (p. 317-339). Oxford: Oxford University Press.

Gerring J. (2010). Causal Mechanisms: Yes, But... *Comparative Political Studies, 43*(11), p. 1499-1526.

Gerring J. (2012). *Social Science Methodology: A Unified Framework*. Cambridge: Cambridge University Press.

Gill J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly, 52*(3), p. 647-654;

Hay C. (2006). Political Ontology. In Robert E. Goodin and Charles Tilly (Eds.) *The Oxford Handbook of Contextual Political Analysis* (pp. 78-96). Oxford: Oxford University Press.

Hoover K., Siegler M. (2008). Sound and Fury: McCloskey and Significance Testing in Economics. *Journal of Economic Methodology, 15*(1), p. 1-37.

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

McCloskey D. (1985). The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests. *American Economic Review, 75*(2), p. 201-205.

Meehl P. (1967). Testing Theories in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*(2), p. 103-115.

Meehl P. (1990). Appraising and Amending Theories: The Strategy of Lakatosian Defence and Two Principles What Warrant It. *Psychological Inquiry, 1*(2), p. 108-141.

Mill J. (1974). *A System of Logic Ratiocinative and Inductive*. Toronto: The University of Toronto Press.

Quine W. (1951). Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review, 60*(1), p. 20-43.

Rainey C. *Testing Hypotheses of No Meaningful Effect.* Retrieved August 23 2014, from http://www.polmeth.wustl.edu/media/Paper/nme.pdf.

Rosato S. (2003). The Flawed Logic of Democratic Peace Theory. *American Political Science Review, 97*(4), p. 585-602.

Ross M. (2001). Does Oil Hinder Democracy? *World Politics, 53,* p. 325-361.

Rozeboom W. (1960). The Fallacy of the Null-Hypothesis Significance Test. *Psychological Bulletin, 57*(5), p. 416-428.

Schrodt P. (2014). Seven Deadly Sins of Contemporary Quantitative Political Analysis. *Journal of Peace Research, 51*(2), p. 287-300.

Stanford K. (2013). *Underdetermination of Scientific Theory.* Retrieved August 18 2014 from Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/scientific-underdetermination/.

Ward M., Greenhill B., Bakke K. (2010). The Perils of Policy by p-value. *Journal of Peace Research, 47*(4), p. 363-375.

Welzel C. (2009). Theories of Democratization. In Ronald F. Inglehart, Christian Haerpfer, Patrick Bernhagen, Christian Welzel (Eds.) *Democratization* (pp. 74-90). Oxford: Oxford University Press.

Welzel C., Inglehart R. (2008). The Role of Ordinary People in Democratization. *Journal of Democracy, 19*(1), p. 126-140.

Western B., Jackman S. (1994). Bayesian Inference for Comparative Research. *The American Political Science Review, 88*(2), p. 412-423.

Zumbo B., Hubley A. (1998). A Note on Misconceptions Concerning Prospective and Retrospective Power. *The Statistician, 47*(2), p. 385-388.

Мельвиль А.Ю., Стукал Д.К. (2011). Условия демократии и пределы демократизации. *Полис (Политические исследования)*, *3*, с. 164-183.

Contact Details and Disclaimer:

Ilya Lokshin

National Research University Higher School of Economics (Moscow, Russia). Political Science Department, Lecturer. Laboratory for Qualitative and Quantitative Analysis of Political Regimes, Researcher.

E-mail: ilokshin@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.