

Evaluating Theory in Political Science*

Abstract

This paper discusses several criteria over which theories should be evaluated. After defining the term 'theory,' and identifying the purpose of theory, I argue that any given theory should be comparatively evaluated against a rival theory that purports to explain the same phenomenon along the following criteria: postdictive accuracy, generality of explanans, hypothetical yield, progressive research program, implications, and parsimony.

August 9, 2001

Will H. Moore
Department of Political Science
The Florida State University
Tallahassee, FL 32306-2230
whmoore@garnet.acns.fsu.edu
<http://garnet.acns.fsu.edu/~whmoore/>

1 Introduction

This essay makes a case for invoking a specific set of criteria for evaluating theories in political science.¹ If we are to make knowledge claims about politics then we need to specify a set of criteria that enables us to evaluate those claims. That is, since political scientists are interested in using science to develop knowledge claims, we need to specify how this is done, and the major task is to identify criteria that enable us to determine which of any two knowledge claims has more value for answering the questions that interest us.

The essay proceeds as follows. First I define the term ‘theory’. Second, I discuss a bogus criterion that some scholars find useful: the veracity of assumptions. Having dispatched that criterion, I turn my attention to a discussion of the criteria political scientists should use to compare rival theories: empirical accuracy of hypotheses, generality of explanans, hypothetical yield, progressive research program, breadth of policy implications, and parsimony.

2 Defining Theory

The term ‘theory’ should be reserved for collections of statements that propose causal explanations of phenomena and meet the following three criteria. First, most political scientists would agree that the statements that compose a theory should be internally consistent. Second, most political scientists would also agree that theories should be logically complete (i.e., the hypotheses deduced from the theory should follow logically from the assumptions of the theory). Third, most political scientists would agree that the set of statements must have falsifiable implications.

King, Keohane & Verba (1994) make a useful distinction between descriptive and causal inference, explaining that causal inferences concern the causal relations between two concepts whereas descriptive inferences seek to describe the empirical world. A causal explanation is one that describes the relationship among concepts in a theory. That is, a causal explanation of judicial appointments (e.g., US Presidential appointments to the Supreme Court) would define a set of concepts (perhaps ‘appointment,’ ‘partisanship,’ ‘judicial record,’ etc.) and then describe the expected relationships

¹I should note that I can see no reason why these criteria should only be applied to political science as opposed to, say, chemistry or economics. I limit my attention to this field only because it is my field (i.e., the one in which I am interested).

among those concepts (perhaps arguing that both ‘partisanship’ and ‘judicial record’ have a positive causal impact on ‘appointment.’) A descriptive explanation, on the other hand, would simply record observations about judicial appointments (e.g., President Bush’s appointees, their partisanship, judicial records, etc.). Put differently, causal explanations are explicitly theoretical whereas descriptive explanations are implicitly theoretical (if one subscribes to a postmodern ontological position) or atheoretical (if one subscribes to a modern ontological position—see Moore (2001)).

A set of statements is internally consistent when the statements do not contradict one another. Put differently, theories should be logical (more precisely, the statements of which a theory is made should be logically consistent with one another). Few political scientists use formal logic to state their theories (using formal logic would make it easier to determine if the statements that compose the theory are logically consistent with one another), but more and more (though still very few) political scientists are using mathematics to state their theories.² Though the formal languages of logic and mathematics make it easier to determine whether the statements that comprise a theory are logically consistent with one another, most theories in political science are simple enough that formalization is not necessary: careful scrutiny can be sufficient.

With respect to the second portion of the definition, a theory is logically complete when the hypotheses one deduces from it flow logically from the assumptions. Zinnes (1976) pointed out that many hypotheses in international relations are *ad hoc*, by which she meant they had not been deduced from a set of assumptions.³ Political scientists often use the term ‘theory’ rather loosely: it is not unusual to find a scholar assigning a name to a posited set of relationships among variables and/or empirical regularities and referring to it as a ‘theory.’ I adopt a more restricted and explicit definition of theory, making use of the distinction that Zinnes drew between collections of internally consistent statements that logically lead to hypotheses and other forms of theoretical argument. Sets of statements that are not logically complete are arguments or pre-theories, but they are not theories.

The third part of the definition, falsifiable implications, concerns (1) circularity and (2) the specificity of hypothetical statements and whether they make claims about the empirical world. Circular (or tautological) sets

²Game theory is becoming an increasingly popular mathematical tool for developing theories, but other types of formal, mathematical models are also used by political scientists. See Morrow (1994) for an introduction to game theory and Rapoport (1983) for an introduction to other mathematical models.

³Morton (1999) refers to these arguments as non-formal theories.

of statements cannot be tested—they are logically valid by definition, but they also tell us nothing about the empirical world, and thus are not theories. Second, theories of politics must have hypothetical implications about politics on our planet, not politics in an imaginary world. Further, the claims must be specific enough that they can be wrong. In other words, they should be specific enough that it is clear what evidence is required to determine whether the evidence is consistent or inconsistent with the implications. That said, we should not require that the instruments that may be required to collect relevant evidence actually exist. That is, the implications of the theory should be falsifiable in principle.⁴

We ought to use terms like ‘argument,’ ‘framework,’ ‘perspective,’ etc. to label explanations that do not meet these criteria.

3 A Bogus Criterion

Some scholars like to consider whether the assumptions of a theory are realistic. That is, they contend that one should prefer theories with realistic assumptions to those with heroic or unrealistic assumptions. Some scholars suggest that a useful criterion for evaluating theories is the veracity of the assumptions they make (e.g., see Moe (1979)). On the other side of the debate, Friedman (1953) provides the classic defense of the argument that the evaluation of theories should focus on the hypotheses not the assumptions. As long as one agrees that the goal of theorizing is to provide heuristic explanation, Friedman is right, and one should not seek to evaluate the assumptions of a theory. I only raise this criterion because some scholars believe it is useful, and since I do not agree, it is important to explain why.⁵

The argument that theories should contain realistic assumptions runs something like this. If theories reveal Truth, then one cannot prefer theories with unrealistic assumptions to theories with realistic assumptions. That is, if the purpose of theory is to reveal Truth, then it does not make much sense to use theories with unrealistic assumptions: falsehood cannot lead to Truth.

However, if theories are understood to provide heuristic explanation,

⁴I am reminded of a story out of the National Center for Atmospheric Research several years ago where scientists had finally been able to lower the temperature within a tube close enough to absolute zero to test the implications of a theory proposed by Einstein back in 1913. Though the implications of the theory could not be tested for eight years, they were still falsifiable in principle.

⁵As explained in Moore (2001), I take a postmodern–pragmatist position, and those who hold a modern–realist position would reject the argument I make here.

then it does not make much sense to evaluate theories by evaluating the validity of the assumptions upon which the theory is built. Instead, it makes more sense to evaluate the hypotheses implied by the theory, as Friedman argued.

I can imagine an objection which might proceed as follows. Imagine a situation where two theories produced precisely the same hypotheses. Imagine further that one theory had assumptions that were unobjectionable on their face, whereas the other had assumptions that were fantastic on their face. Under such circumstances, a critic might submit that one should prefer the first theory to the second.

The idea of assessing the ‘face validity’ of assumptions will hopefully make readers uncomfortable—I am uncomfortable with it. The response to this objection is to inquire how one would judge face validity, which refers to the extent to which the assumptions seem consistent with the present body of knowledge. ‘The present body of knowledge’ is not Truth, but rather a research program that may or may not have rivals, may or may not be progressive (see below), etc. Thus, the face validity of assumptions criterion is intimately tied to a specific research program. As such, it cannot be used as a criterion on which to evaluate the research program. The upshot of this argument is that the ‘validity of the assumptions’ is not a useful criterion for evaluating theories.

4 Theory Evaluation: Some Criteria for Comparison

I describe six criteria over which political scientists should compare theories. First, many scholars argue that theories should be evaluated based on their ability to produce hypotheses that are consistent with relevant evidence. Second, many scholars contend that one should prefer general theories to less general theories. Third, one should prefer theories that produce several hypotheses to those that produce few. Fourth, Lakatos (1970) has proposed that we evaluate research programs rather than individual theories, and that we do so on the basis of whether they are degenerative or progressive. Fifth, we might compare the implications, preferring theories that have several policy implications to those which have few. Finally, parsimony or simplicity is generally considered a virtue.

4.1 Postdiction

Postdiction is the effort to explain past events. Causal hypotheses produce expectations about causal relations among variables. We can test those expectations by treating them as postdictions and comparing them against relevant evidence. When the evidence is consistent with the hypothesis, then the theory is considered useful for explaining the phenomenon under investigation. When it is not, then the theory—as presently constituted—is not considered useful for explaining the phenomenon under investigation.

This discussion requires a couple of clarifications. Lakatos (1970) rejected Popper's (1968) naive falsificationism. Popper argued that hypothesis testing serves the purpose of ruling out theories even though it cannot demonstrate that a theory is 'true.' That is, Popper urged scientists to test the implications of their theories by comparing the implications with relevant evidence. If the evidence is inconsistent with the implications, then Popper argued one can reject the theory. However, if the evidence is consistent with the implications from the theory, Popper argued that we cannot accept the theory for we may later find different evidence that is inconsistent with the theory (perhaps with a different implication). Thus, we tentatively accept the theory and continue testing.

Lakatos (1970) labeled Popper's argument 'naive falsificationism' because Popper did not appreciate that we cannot use evidence to reject a theory any more than we can use it to accept a theory. Lakatos' argument suggests that we may choose not to use the theory to account for that phenomenon, but most theories yield several hypotheses across different phenomena, so it does not make much sense to toss a theory on the scrap heap because of a single poor showing. In addition, revision to the theory may change the hypotheses such that they now postdict well. Yet a third reason to avoid naive falsificationism is that the empirical test may have been faulty. All such tests are exposed to experimental error, measurement error, poor design, etc. Thus, a single test may prove to have been the victim of some sort of error, and we should be reticent to draw strong conclusions from any individual test. Lakatos' views are discussed in more detail below.

That said, postdiction is nevertheless an important criterion. I contend that it is the first criterion one should consider when comparing two theories. If, with respect to the phenomenon under investigation, they produce distinct hypotheses, then one should determine the extent to which each is consistent with the relevant evidence. Keeping Lakatos' warnings against naive falsificationism in mind, one should provisionally prefer the theory

that postdicts well to the one that does not. If the theories produce the same hypothetical expectations with respect to the phenomenon under investigation, then the secondary criteria discussed here should be considered.

4.2 Generality

The next criterion to consider is generality. This criterion concerns the breadth of the scope of the explanans (i.e., the phenomenon we are trying to explain). That is, ‘How many different phenomena can the theory account for?’ In other words, ‘How broad, or general, is the explanans of the theory?’ The answer to this question hinges on the generality or breadth of conceptualization: where theory *A* may seek to explain all human interaction, theory *B* may seek to explain all human interaction characterized by attributes *X* and *Y*. In that case, theory *A* is more general than theory *B* and it is to be preferred in the circumstance where both theories produce the same hypotheses regarding a given research question.

As an example, Skocpol (1979) proposes a theory of social revolution. She notes that the explanans is quite limited: there have been fewer than perhaps half a dozen such events. Further, she criticizes scholars such as Gurr (1970) who propose theories of general concepts such as civil strife, arguing that they are lumping together rather distinct phenomena. A theory of violence (i.e., all violent human behavior) is more general than a theory of social revolution. Further, a theory of social revolution is more general than a theory of the French Revolution. We prefer theories that are more general than their competitors because they are more powerful—with them we can explain more events than we can with a less general theory.

4.3 Hypothetical Yield

Another useful criterion concerns the number of hypotheses the theory yields beyond the issue under investigation. Assume that theory *A* and theory *B* each produce the same hypotheses with respect to the topic of investigation. We invoke the hypothetical yield criterion when we ask what other hypotheses the theory yields. If theory *A* yields additional hypotheses that address other topics, and theory *B* does not, then one should prefer theory *A*.⁶

Like generality, hypothetical yield concerns power. The more hypotheses a theory produces, the more phenomena it can account for, and more is better!

⁶I am assuming that the additional hypotheses are consistent with relevant evidence.

4.4 Degenerative versus Progressive Research Programs

As noted above, Lakatos (1970) proposes a set of criteria for evaluating rival research programs. In brief, Lakatos agrees with Popper's (1968) argument that the problem of theoretical over determination (among other problems) prevents scientists from using hypothesis tests to adjudicate among rival theories. The problem of theoretical over determination is the problem that any given hypothesis can be produced by more than one theory (and here we include theories that humans have yet to invent). Because no given hypothesis is uniquely implied by a single theory, evidence that is consistent with a given hypothesis does not enable us to accept that theory. However, Lakatos takes it further, arguing that not only is hypothesis testing incapable of establishing the validity of a theory, it is also incapable of establishing the invalidity of a theory. Lakatos argued that while a given hypothesis can be invalidated by empirical evidence, the theory that produced it cannot be so impugned. The reason is that theories produce many hypotheses and should not live or die on the basis of a test of a single hypothesis. Rather, theories should compete with one another.⁷ The idea is that if a theory produces several hypotheses, and two are consistent with the evidence, but the others are not, then we ought to retain the theory unless there is a rival theory that can also account for the two phenomena that our theory can account for. Stated simply, the dictum is: do not throw out useful theories, even if they are only partially useful.

Lakatos expected scientists to revise their theories over time. Sometimes those revisions might change the hypotheses implied by the theory (for example, a theory might be made more specific or it might be made more general). If the old hypotheses are no longer implied after revisions, and the new hypotheses are consistent with relevant evidence, then the theory is more useful than it used to be. However, Lakatos does not consider all revisions equally useful. He distinguishes between two types of revisions, and to understand this, it is useful to turn our attention to his discussion of research programs.

For Lakatos, evidence is not used to falsify hypotheses and, thereby, impugn the theories that produced them. It is reasonable to consider the hypothesis falsified, but theories form research programs, and research programs do not live and die as a consequence of hypothesis tests. Rather, research programs are best evaluated by considering the changes made to the theory when it is faced with contrary evidence. Lakatos proposes that

⁷Actually, Lakatos argued that research programs should compete with one another, but we will discuss theories for the time being.

scholars distinguish between progressive and degenerative research programs to evaluate revisions made to theories.

The distinction between progressive and degenerative hinges largely on the domain of the explanans (or, if you prefer, the theory's generality) and the restrictions placed on the theory by auxiliary assumptions. A progressive research program expands its explanans, a degenerative one reduces (or, over a period of time, fails to expand) its explanans. A progressive research program eliminates restrictive auxiliary assumptions, a degenerative one adds them. Contrary evidence challenges research programs in that it provides a motive to either (1) reduce the explanans or (2) add assumptions in an effort to insulate the theory from the contrary evidence. Thus, to evaluate a research program, one needs to determine whether it is progressive or degenerative, and one does so by determining whether the explanans covered by individual theories has been expanded or reduced, and whether restrictive auxiliary hypotheses have been added or eliminated.

4.5 Implications

Implications is, perhaps, an unusual criterion. I have in mind comparing theories with respect to the policy implications that they yield. The motivation behind proposing this criterion is the contention that theory should help us understand the world so that we can act in the world. Thus, one should prefer a theory that, *ceteris paribus*, yields several policy implications to one that yields few.

4.6 Parsimony

Parsimony is a criterion that is invoked in situations where the postdiction criterion cannot separate two theories. It follows from Ockham's razor which suggests that theories should be simplified—unnecessary assumptions should be pared. If theory *A* invokes two assumptions and contains three terms while theory *B* invokes six assumptions and contains six terms, then theory *A* is more parsimonious than theory *B* and one should, *ceteris paribus*, prefer theory *A*.

5 A Hierarchical Set of Criteria

A brief recapitulation of the criteria and the hierarchy may prove useful. In the preceding section I identified six criteria we can use to comparatively evaluate theories and research programs. Yet, before we can compare two

theories, we must determine whether they are in fact theories or if they are what I have called theoretical arguments. Theories are sets of statements that specify causal relations among the concepts they contain, and those statements must be internally consistent, logically complete, and have implications for the empirical world. Once we have two theories that both purport to explain the same phenomenon, then we can concern ourselves with the six criteria. The first criterion to consider is postdiction. As such, the first task when evaluating theories is to determine whether they postdict well. If the two theories can be separated on their ability to postdict, no reference is made to the remaining criteria.⁸ If the theories postdict the phenomenon of interest equally well, then to determine which is the more useful theory, one might determine if they imply distinct hypotheses about a different phenomenon, and test those hypotheses. While such a critical test is useful, one should consider the remaining criteria as well.

If the theories cannot be separated on postdiction, then the political scientist should turn to the remaining criteria: generality; hypothetical yield; degenerative versus progressive research programs; parsimony; and implications. In a case where two rival theories postdict equally well⁹ the political scientist should turn her/his attention first to generality, then to hypothetical yield, etc. and continue down the list until the theories can be separated.

References

- Friedman, Milton. 1953. The Methodology of Positive Economics. In *Essays in Positive Economics*, ed. Milton Friedman. Chicago: University of Chicago Press pp. 3–43.
- Gurr, Ted Robert. 1970. *Why Men Rebel*. Princeton: Princeton University Press.
- King, Gary, Robert O. Keohane & Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
- Lakatos, Imre. 1970. Falsification and the Methodology of Scientific Research Programmes. In *Criticism and the Growth of Knowledge*, ed. Imre Lakatos & Allan Musgrave. New York: Cambridge University Press pp. 91–195.

⁸As an exercise, you might ask yourself what Lakatos would have to say about this sentence.

⁹Please note that this is unlikely.

- Moe, Terry M. 1979. "On the Scientific Status of Rational Choice Theory." *American Journal of Political Science* 23:215–243.
- Moore, Will H. 2001. "Observing the Political World: Ontology, Truth, and Science." Florida State University.
- Morrow, James D. 1994. *Game Theory for Political Scientists*. Princeton: Princeton University Press.
- Morton, Rebecca B. 1999. *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. New York: Cambridge University Press.
- Popper, Karl R. 1968. *The Logic of Scientific Discovery*. New York: Harper Torchbooks.
- Rapoport, Anatol. 1983. *Mathematical Models in the Social and Behavioral Sciences*. New York: Wiley.
- Skocpol, Theda. 1979. *States and Social Revolutions*. New York: Cambridge University Press.
- Zinnes, Dina A. 1976. *Contemporary International Relations Theory*. New York: The Free Press.