Validating and Verifying Validation and Verification:
The Methodological Challenge of a Public Policy Imperative

Ian S. Lustick

University of Pennsylvania

Abstract

Verification and Validation are legally required evaluations that the Defense Department must make of any system, model, or simulation it deploys. With extensive experience and detailed protocols for verifying and validating natural science based systems, DoD now faces the problem of how to conduct such evaluations for the social science based systems, models, and simulations it is increasingly interested in acquiring. Employing an "epistemological ladder" or hierarchy linking ontology to data, this paper will analyze the verification and validation process as entailing analogous questions about warrantability of inferences and accuracy or generalizability of findings appropriate at any step in the ladder. Analysis of the confusing impact of the "two cultures" idea and its importance within influential philosophy of science texts, along with the particular history of validation and verification concepts within the American military, will be used as the basis for presenting a unified approach to the problem. The usefulness of the approach is tested in a mapping exercise of key debates between quantitatively and qualitatively oriented political science methodologists and of the conventional wisdom reflected in recent DoD documents.

Paper prepared for presentation at the annual meeting of the American Political Science Association, New Orleans, LA, August 30-September 2, 2012. Support for this research came from the Lockheed Martin, Advanced Technology Laboratory's Model Evaluation, Selection, and Application program, the Office of Naval Research, and the Bess W. Heyman Chair, University of Pennsylvania.

Model Validation and Verification as Problematic Imperatives

for the Department of Defense

The Department of Defense (DoD) fights wars and prepares to fight wars. When it makes mistakes or uses a weapons system or a support system that performs inadequately, the results are painful. The costs are measured not only in money but in lives. Accordingly, DoD invests significant resources in pre-deployment assessments of the systems, models, and simulations it purchases. These investments are based on a well-corroborated judgment that significant expenditure on such "Validation and Verification" is a cost-effective way to reduce error and error-produced costs and casualties. Indeed, no weapons system is eligible for deployment by DoD unless it has been "validated, verified, and accredited." As with virtually everything else in the military, the process has been acronomyized: VV&A.[1]

It is difficult to know how meticulously this process is followed within DoD and how much variation exists in the application of its protocols. Whatever flexibility or incompleteness the process may have, and however rough the rules may be for deciding how rigorously the process should be enforced in any domain or for any particular technology, there is a general view that a new and distinctive set of protocols is needed for carrying out VV&A for social science based systems, models, and/or simulations. There has been rapid increase in the military's dependence on such models for operating in Afghanistan, Iraq, and other "battle-spaces" featuring complex social, cultural, economic, and political factors. This dependence coupled with the practical and legal requirement for validation and verification has produced confusion about how to assess social science models or social science-based systems. It is in this context that the Office of Naval Research issued a Broad Agency Announcement (BAA) soliciting research and development proposals for devising best practices and theoretically grounded protocols for VV&A as applied to "non-kinetic" systems—systems intended for deployment in connection with understanding, analyzing, forecasting, and/or manipulating human societies or social formations.

The BAA itself implies a distinction between the natural or physical sciences, and the social sciences. Were no differences of kind imagined to exist, there would be no need for a new or distinctive set of protocols for VV&A regarding social science based models. Contrary to the analysis implicit in the BAA, this paper advances the argument that to extend V and V to social science based models will require understanding the deep sense in which they are *not* different from natural science based models. There is one scientific method; responding to one generic kind of problem—how to wrest understanding of the world from patterns of behavior that are often difficult to discern, awkward to study, and uncertainly documented. The implication of the unity of the scientific method is that however different is the level and sophistication of our scientific knowledge in a particular domain, the requirements for evaluating science-based models are fundamentally equivalent in every basic respect regardless of whether the domains are wholly or partially affected by human behavior or whether human behavior is absent entirely.

I begin with a close look at the exact wording of the Defense Department's 2009 solicitation for proposals to address the problem of VV&A for social science models. This is followed by a brief discussion of the traditional "two cultures" distinction, how it is expressed in two foundational scientific methodology texts—Thomas Kuhn's *The Structure of Scientific*

Revolutions and Imre Lakatos's essay on the methodology of research programs, and how it has shaped the point of view reflected in this BAA. After advancing the argument that this distinction is largely unjustified, I will present a schema for organizing thinking about and generating guidelines for Validation and Verification across the "two cultures" divide based on an analysis of the logical requirements of any scientific model. After using the concept of "construct validity" to clarify the relationship between "verification" and "validation" I will use this organizing framework to map patterns in contemporary methodological disputes among social scientists to the dilemmas faced by the Department of Defense in seeking V and V solutions for social science models. I will conclude by suggesting that the advice contained in the BAA, highlighting the importance of attending to the quality and rigor of theories, is justifiable, supported by an important part of the critique leveled by qualitative researchers against dominant quantitatively oriented methodological norms, but insufficient for satisfying the VV&A requirements that social science, as well as natural science, based models should be required to meet.

In 2009 the Office of Naval Research published a BAA soliciting proposals for strengthening V and V capabilities with respect to Human, Social, Cultural, and Behavioral (HSCB) models.

> *System to assess and select socio-cultural behavior models:*
>
> **Models of human socio-cultural behavior cannot be validated** in the traditional military sense. Because of the deep uncertainty associated with the variables that are the focus of these models, they cannot reliably be validated ***by traditional means (i.e., against physical or historical reality).*** Instead, they must be assessed for the quality and rigor of their theoretical underpinnings and the integrity of the models' construction. Consequently, putting HSCB models into operational usage will require a regular standardized assessment procedure supported by structured data and information systems. We seek projects that will develop and demonstrate an integrated set of model description data (metadata), information systems, and procedures that will facilitate assessing the software engineering quality of sociocultural behavior models, their theoretical foundation and the translation of theory into model constructs. Such a system should facilitate assessing the conceptual, computational and theoretical relationships expressed both in isolated models, and for a system of models. In addition, the metadata should also allow users' selection of models based on "granularity" (national, regional, group-level or individual), and the time horizon of projections (hours … years). The metadata should also incorporate a taxonomy of model-types (e.g., causal, event-driven, agent-based, descriptive, statistical, etc.).[2]

The phrases I have bolded in this quotation state clearly the fundamental distinction being made between VV&A for natural science based models, and those the authors of the BAA anticipate as appropriate for social science based models. It is important to notice what, "instead" of traditional methods, is to be done. By emphasizing assessment of the theoretical underpinnings of these models and the "integrity of the models' construction," the BAA would seem to be advising, or even mandating, virtual abandonment of the idea of "validating" HSCB models, in the sense of measuring the degree to which they actually do the work in the world

they purport to do or were designed to do. This abandonment is justified on the basis of what seems to be a fundamental distinction between natural science and social science.

In contrast to the natural science based models that the military "traditionally" subjects to validation and verification, the variables in social science models are described as "associated" with "deep uncertainty." To be sure, the authors of the BAA do suggest that a particular kind of validation is possible with social science models; but only assessments focused on the value of the theories the model is putatively operationalizing—those that "underpin" it. By characterizing validation for social science models as *only* requiring evaluation of the theoretical provenance of such models, along with the fidelity of their construction to their theoretical underpinnings, and *not* requiring evaluation of their performance via comparisons with "physical or historical reality," one of two propositions is implied. Either natural science variables are in some qualitative fashion more precisely conceptualized and operationalized than social science variables, or social science theories describing relationships among variables have very little corroboration compared to natural science theories.

The authors of the BAA cite an additional rationale for loosening requirements for V and V with respect to social science-based models. While "physical or historical reality" can be used to test the claims of natural science based models; they declare that the performance of social science-based models simply cannot be measured against empirical evidence. This is a drastic judgment and would seem, by ruling out efforts to determine whether a social science based model actually "worked," to render V and V as required by the DoD of all models and systems it uses, impossible. At minimum, this call for proposals seems to lower the bar for V and V with respect to social science based models by in effect focusing on "verification" and by investigating the "validity" of models based only on the reputation and power of the theories operationalized by them, rather than by measuring the effectiveness of their output for meeting intended purposes.

On the other hand, the entire purpose of the BAA was to fund efforts to develop ways to do just that. Clearly there is a tension here, between legal, organizational, and budgetary imperatives to treat social science models as any other kind of scientific model is to be treated (*i.e.* to subject it to VV&A requirements) and an image of social science and the models that arise for understanding HSCB phenomena as too underspecified and too weak to justify fundamental evaluative questions entailed by VV&A. In the analysis that follows I shall seek to resolve this tension by clarifying the meaning of verification as a form of validation, explaining the non-analytic sources of the belief that social science models are qualitatively different from natural science models, and mapping key methodological debates in the social sciences (in political science in particular) to the problems with which the authors of the BAA are wrestling.

The Unity of Science Despite Prevalent Beliefs to the Contrary

In 1950 C.P. Snow's Rede Lecture at Cambridge University, *Two cultures and the Scientific Revolution*, created a meme and expressed an abiding disposition among natural scientists when he declared that intellectual life had been bifurcated between "men of science" and "literary intellectuals." Bemoaning what he considered the outsized influence on public thinking and public policy exercised by the latter, Snow later urged scientists to speak directly to the public so that the rationality and systematic thinking associated with scientific methods and

knowledge could be exploited for the general welfare. Snow's image was of a virtually unbridgeable gulf between those whose work could be expressed mathematically, and those—"the literary intellectuals"—exercised an overly powerful influence over public policy and whose virtuosity with discursive expression disguised their profound ignorance of science and technology. In this lecture, Snow altogether ignored the "social sciences." When he referred to science, his audience could be sure he meant biology, physics, chemistry, or astronomy, not sociology, anthropology, or political science.

Snow's claims sparked an intense and wide-ranging debate. Its vituperativeness seemed at once to confirm his judgment and even widen the gap between the two communities, as he constructed them. Several after the lecture, Snow felt compelled to revisit his analysis. Without ever using the term "social science," he suggested in this updated analysis that a new "body of intellectual opinion" was "forming itself, without organization, without any kind of lead or conscious direction." Here, he opined, a "third culture" was perhaps emerging.

> This body of opinion seems to come from intellectual persons in a variety of fields—social history, sociology, demography, political science, economics, government (in the American academic sense), psychology, medicine, and social arts such as architecture. It seems a mixed bag: but there is an inner consistency. All of them are concerned with how human beings are living or have lived—and concerned, not in terms of legend, but of fact. I am not implying that they agree with each other, but in their approach to cardinal problems…they display, at the least, a family resemblance.[3]

For my purposes what is striking about Snow's updated treatment is that even as he nods toward the existence of a "fact"-oriented approach to the understanding of human behavior, he either cannot bring himself to label the disciplines involved as "scientific." Most likely, he was so completely captured by the common sense of his ordinary language, that nothing that was not familiar to him through the physical or natural sciences could be labeled or recognized as "science." Instead Snow suggests that whatever value there might be in these approaches to the "facts" of human behavior and experience, the overall approach should be considered a "third culture." It would not be the discourse of "literary intellectuals," but neither would it be "science."

However, this approach left Snow bereft of a solution to the question of how work in this area could be judged. Since its purpose was to ascertain and generalize from "facts," rather than discuss meanings or symbols in "legends," a replicable method and some rules for the use of evidence would be required. But Snow did not suggest what method this might be. Had he sought to answer that question it would have been difficult for him to avoid the conclusion that the "scientific method" would be appropriately, even necessarily, applied.

One of Snow's core ideas, that the natural sciences have constituted the only domain within which systematic empirical work can contribute to theoretical knowledge, finds more pointed, and usually less polite, expression in the work of influential philosophers and historians of science. This point can be made clearly by considering how much difficulty Thomas Kuhn and Imre Lakatos, two of the most influential analysts of both the history and methodology of science, have with respect to classifying the activities of social scientists as "science," that is as

being engaged as "scientists" in a knowledge producing, knowledge accumulating, process of devising theories, comparing them to empirical data, and revising those theories in light of data. In the introduction to his seminal volume, *The Structure of Scientific Revolutions*, Kuhn attributes the inspiration for writing the book to the contrast he noted between the orderly, systematic, and progress producing conversations among natural scientists (guided and limited by paradigms of agreed upon knowledge and procedure) with the "controversies over fundamentals" and the "number and extent of overt disagreements between social scientists about the nature of legitimate scientific problems and methods."[4] Kuhn's entire presentation of the history of "science" via paradigms, normal science within paradigms, and displacement of paradigms by others, is comprised only and entirely of examples drawn from the natural or physical sciences—biology, astronomy, electro-magnetism, physics, and chemistry. Often in the text Kuhn uses "physical scientists" as a synonym for "scientists."[5] In his one explicit reference to politics, and by implication to political science, Kuhn refers to the "vast and essential differences between political and scientific development." [6]

Yet Kuhn's very argument makes it impossible for him completely and explicitly to classify the study of human behavior and of human societies as "unscientific." Repeatedly he draws on experiments performed by psychologists demonstrating the powerful constructivist effect of expectations and a prior beliefs on perceptions. Kuhn uses these findings to support his argument that scientists operating within different paradigms effectively live in, and thus see, different worlds—the deep presumptions about the world that led scientists to see electricity as a fluid are treated as analogous to the deep presumptions about a deck of cards that lead subjects to be unable to "see" a card presented as a "red spade" or a "black diamond." Without ever using the history of psychology as data for testing his theory of scientific development, he does, in other words, use the products of that science, treated as valid, as part of his effort to validate his theory of science in general. Even more striking is Kuhn's embrace of a purely political concept, namely "revolution," as the theoretical fulcrum of his entire approach. Kuhn does note the irony of his use of a concept associated with politics for analyzing science, despite the "vast and essential differences" he believes separate these realms. Yet he neither acknowledges nor analyzes the implications of his use of a detailed model of revolution to produce specific hypotheses about the phases of paradigm change and the types of scientists likely to engage in the challenge of old paradigms and the development of new ones. Nor does Kuhn acknowledge the contributions of political scientists to the production of the theories of revolution which, in condensed form, he is effectively applying.

Imre Lakatos is best known for his integration of Popperian falsificationism with Kuhnian patterns of contestation among paradigms—termed "research programs" by Lakatos. His work has been enormously influential among social scientists, and among political scientists in particular, in spite of the vehemence of his dismissal of even the kind of social science—psychology—that Kuhn explicitly drew upon for support of his theory. As is the case with Kuhn, Lakatos developed, presented, and defended his theory of what science is and how it progresses by drawing exclusively on the history of disciplines within the natural sciences. In his "ordinary language," Lakatos uses the terms "science" and "natural science" to mean the same thing. The only explicit references he makes to the study of human behavior are to social psychology and the sociology or psychology of knowledge. His substantive point is that they do not qualify as science since they do not meet his definitional requirement, *viz.* "*continuous*

growth." In any event Lakatos's sarcastic dismissal of work in such fields illuminates even more clearly how aligned he is with the Snow/Kuhn disposition to consider claims of a "social science" to be at best descriptors of an "immature" science and at worst, pure charlatanism. His definition science, he says,

> hits patched-up, unimaginative series of pedestrian 'empirical' adjustments which are so frequent, for instance, in modern social psychology. Such adjustments may, with the help of so-called 'statistical techniques', make some 'novel' predictions and may even conjure up some irrelevant grains of truth in them. But this theorizing has no unifying idea, no heuristic, power, no continuity. They do not add up to a genuine research programme and are, on the whole, worthless.[7]

Despite Lakatos's utter disdain for social science, and his refusal to categorize them as eligible for analysis according to his theory of evolving research programs, his analysis is virtually saturated with political models and theories: theories of institutionalization, interest aggregation, and resource mobilization.[8] Indeed, according to Alan Musgrave, Lakatos's long-time collaborator, Lakatos eventually abandoned almost entirely his attempt to explain scientific progress by processes of survival of the methodologically "fittest" and highest-performing theories, embracing instead an approach which explained patterns of scientific change in politically, determined by the relative abilities of scientist-protagonists to mobilize economic, reputational, and institutional resources, both inside and outside the academy.[9]

Within the Department of Defense and the communities of contractors that participate in its evaluative culture, the two cultures outlook, based on a categorical distinction between science—figured commonsensically as the natural or physical sciences--and all other forms of knowledge production, predominates. However, as I have suggested, it has been impossible to achieve a clean and complete distinction between the methods and evaluative procedures appropriate for natural science and those appropriate for the study of human behavior. Although no room is made explicitly in the theories of Snow, Kuhn, Lakatos, for validation and verification tests applied within a social science framework, each presents his ideas and arguments in ways that require them to make social science claims or that imply the appropriateness of assessing the credibility of such claims. My argument is that neither DoD nor the community of experts that services its needs will be able to do what philosophy of science cannot—maintain a consistent evaluative stance with respect to scientific activity that justifies a qualitative difference between procedures for validation and verification across the so-called "two cultures" divide.

Let us consider treatment of this problem by an influential study, *Behavioral Modeling and Simulation: From Individuals to Societies*, commissioned by the Department of the Air Force and the National Academy of Sciences.[10] The committee responsible for this report was headed by co-chairs, Greg Zacharias and Jean MacMillan. Other members of the committee include leading academics with specializations pertaining to psychology, computer simulation, agent-based modeling, and soft-ware engineering. The overall perspective of the authors on validation and verification is in some ways opposite to that of Snow, Kuhn, and Lakatos. Zacharias, *et. al.* believe in social science, and in the policy value of computational and other models based on social science. This orientation is unsurprising given that they are producers of

computational models, simulations, and other social science based products for use by DoD. Indeed the two co-chairs, at the time of their appointment, were leaders of prominent defense consulting firms (Charles River Analytics and Aptima, respectively) specializing, in part, in simulation and social science modeling.

On the other hand, the authors of the report share enough of the two culture approach to argue that standards, rules, and expectations for validation and verification of social science based models cannot be and should not be the same as, or as strict as, those for natural science-based models. The authors cited a 2003 Defense Modeling and Simulation Office conference dedicated to assessing qualitative and quantitative challenges entailed in shifting modeling targets to large groups and societies. In keeping with their general observation of a marked imbalance in attention to V and V issues in domains involving individual, organizational, or societal models, compared with those involving physical science or engineering models, the authors observe that "only a very small fraction of the proceedings directly addressed critical VV&A issues for… behavioral models."[11] As a high profile attempt in DoD to apply V and V standards to large scale social simulations, the authors mentioned the Joint Forces Command's Urban Resolve Program. This involved, *inter alia*, an ambitious simulation of the city of Jakarta. According to Zacharias *et. al.*, however, "it is unclear whether V&V efforts for Urban Resolve went any further than the "looks ok" test. This is not atypical of large-scale simulations in general."[12]

The Zacharias report did make recommendations for validation of social science based models. Although regularly referring to "verification and validation," the report acknowledges its emphasis is on validation (i.e. the accuracy of models) rather than verification (*i.e.* the conformity of the model to design specifications).[13] However, by judging that "universal rules about what is the appropriate procedure for validating IOS models are not possible," the authors effectively rejected the idea of applying to these models the same rules used for V and V with respect to physical science based models. Instead they recommended that validation of "individual, organizational, and societal" models should involve three techniques:

> (1) participation by multiple experts who can provide different perspectives on the action domain, the scenarios, and the if-then rules incorporated in the model;
> (2) docking of similar computational models against one another;
> (3) comparison to qualitative and theoretical studies and previous quantitative results and exploratory testing for a range of outcomes.[14]

Indeed these rules may be the basis for increasing user beliefs in the value of a model, but neither individually nor collectively are they techniques for measuring the absolute or relative effectiveness of a model by comparing its performance, forecasts, or analyses to real world data or replicable and authoritative renderings of real world data. Instead, the authors of this report suggest that reactions by SME's can, in some unspecified way, be utilized for establishing model credibility. Additionally, they suggest, that if the model's inputs, outputs, or categories can be mapped onto and/or interfaced with the inputs, outputs, or categories of other models (whose validity is itself unevaluated), the credibility of the model under evaluation can be enhanced.

The language of the BAA accepted the argument of this study, *viz.* that the absence of consensual social science theories and the difficulties of testing models in complex and

uncontrollable settings at different levels of human social organization, render standard DoD V and V practices inadequate.[15] On the other hand, by challenging performers to devise "a regular standardized assessment procedure supported by structured data and information systems" for the verification and validation of social science-based models, DoD is asking for more than the authors of this report believed was possible or appropriate.

A Validation and Verification Model that Honors the Demands of the Scientific Method

The strategy suggested here for responding to this challenge recognizes that some domains may be more tractable to the production of effective models than others. In addition, it assumes that, regardless of the simplicity, complexity, or "tractability" of any particular domain, processes of systematic knowledge production will encounter the same kinds of challenges. Therefore they can be subjected to the same set of evaluative criteria. In other words, the key to developing standardized assessment procedures for the social sciences is to understand well-theorized natural sciences and the models derived from them as a "special case" within the much more general problem of assessing knowledge production, models, and simulations of problems regardless of how well-corroborated we now believe available theories to be, or how complex we now believe are the domains within which those theories are applied.

The more or less consensus view among philosophers of science seeking a workable neopositivist epistemology for doing science is that simple "falsificationism" is not a serviceable position because no hypothesis can be falsified by inconsistency with experimental outcomes. The reason for this is that causes for the apparent failure of the hypothesis might have been a flaw inherited from the model that generated the hypothesis, the theory that justified the model, the paradigm that enabled the theory, or even the ontology that serves as the metaphysic for the paradigm. Indeed, the "naïve falsificationist's" intention to be ruthless in elimination of hypotheses, models or theories by using experimental data as a guillotine to eliminate false propositions runs into a problem in the other direction on the scientific ladder of abstraction. For the experimenter must not only be unsure of whether his hypothesis, model, or theory (whatever status the "conjecture" is given) is faulty, but also whether the experimental and measurement apparatus, or even the theories and models that are expressed in and implied by the operation of the experimental and measurement apparatus, are flawed.

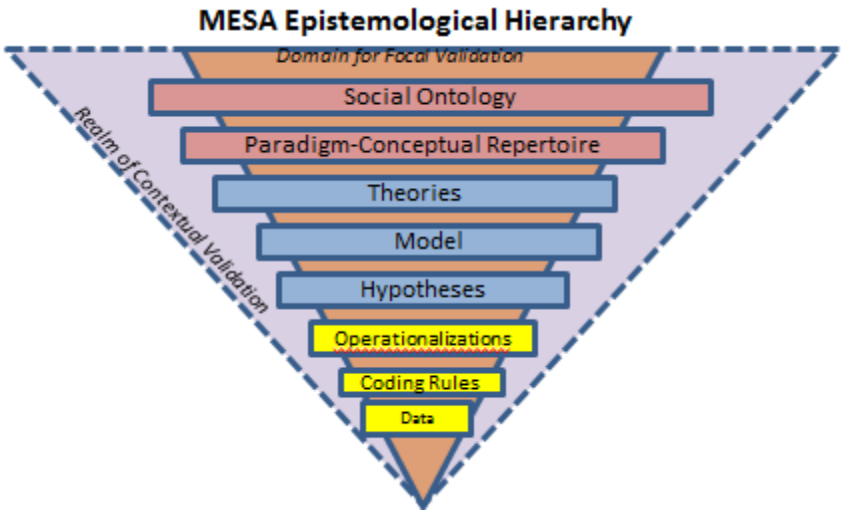
Leaving aside the implication of this analysis for the need to shift attention to "research programs" or their analytic equivalents as units of analysis and as ways to explain scientific progress, we may usefully note a key implication of this now standard critique of positivism for purposes of validation and verification. Just because experimentation yields what may appear to be evidence falsifying a hypothesis, there cannot thereby be an absolute justification for rejecting the hypothesis. For it may be the case that the hypothesis is itself correct, but errors in higher level theories or in theories embedded within measurement instruments are responsible for the result. Likewise, scientists can never be justified in imagining higher-order theories as firmly and absolutely established fountainheads for lower level models or hypotheses whose fate does not and cannot implicate the status of those higher level theories. Social scientists, lacking as much consensus on their own theories and paradigms as is the norm in most physical sciences, are accustomed to this situation, specifically to the implication that a scientist must always be alive to higher order problems that might explain puzzling outcomes of "normal" scientific

inquiry. Many natural scientists may be less accustomed to thinking this way; especially those involved in engineering or computer science, i.e. in the application of knowledge assumed to be well-established and unproblematically corroborated. But this is a "cultural" difference across scientific domains, not a difference springing from any fundamental difference in the epistemological status of tasks undertaken when evaluating theoretically produced hypotheses or inferring theoretical propositions from lower level patterns. Scientific inquiry may be directed at particles, or, just as appropriately, at "particles that think," i.e. theories and models may target material/physical or human/social phenomena. The key point is that in either case, identifying weaknesses, inadequacies, invalid inferences, or inaccuracies associated with a scientific model is an enterprise that must include opportunities to identify such pathologies at any level of an "epistemological hierarchy," from ontology to the coding rules for data collection.

Faced with the public policy task of evaluating the credibility of any particular model, and knowing that shortcomings may be produced by pathologies located anywhere from ontology to data gathering instruments, assessors must have a tool permitting them to parse possible locations of validation or verification errors in distinctive but systematically related levels of analysis. It is also the case that what is treated as a "model" can appear as a generator of very concrete *or* very abstract claims. For example, a model for skill or aptitude testing may specify concrete techniques for gathering and processing information. A model for translating "strain" in a system into different forms of collective behavior will operate at a much more abstract level. For both of these reasons, standardized protocols for evaluating the credibility of models must allow for analogous operations of verification and validation to be performed at multiple points on an "epistemological hierarchy."

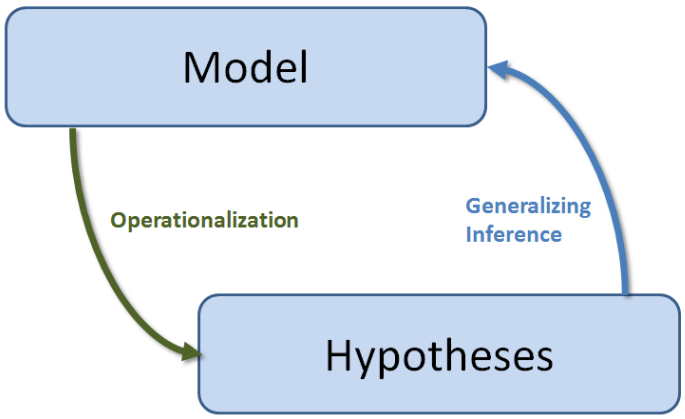
For this purpose I developed, with support from other members of the Lockheed Martin ATL MESA team, a ladder of locations ranging from extremely abstract to extremely concrete. One core scientific enterprise is to operationalize general propositions believed to be true as more specific claims more exposed to testing. Another is to infer more general propositions from specific claims believed to be true. If we consider the relationship between a "model" and a "hypothesis" as generic, that analogous to the relationship between a a theory and a model, or between any general proposition and more concrete "operationalized" versions of that proposition, then we can imagine an epistemological hierarchy or ladder connecting ontological beliefs at the top and observations or data rendered from stimuli at the bottom. At any one level, the higher adjacent rung in this ladder can be thought of as a "model" which is operationalized by "hypotheses" on the rung below it. Simultaneously, that "model," can be thought of as inferred from the performance of "hypotheses" on the adjacent rung below it. Our "model" of this epistemological hierarchy is presented graphically in Figure 1.

Figure 1



One enterprise of science is to look down from any rung and to operationalize a general belief in the form of more concrete claims; for example to operationalize a theory via specific models, or to infer from a model in which one has confidence a number of specific hypotheses worthy of testing. Another scientific enterprise is to look up from any rung and infer more general claims about the world based on a pattern of evidence accumulated from below. Thus a general theory could be inferred from patterns of performance by multiple models, or a model could be inferred from patterns of corroborated hypotheses. These two directions of scientific conjecture and inquiry are illustrated in Figure 2.

Figure 2



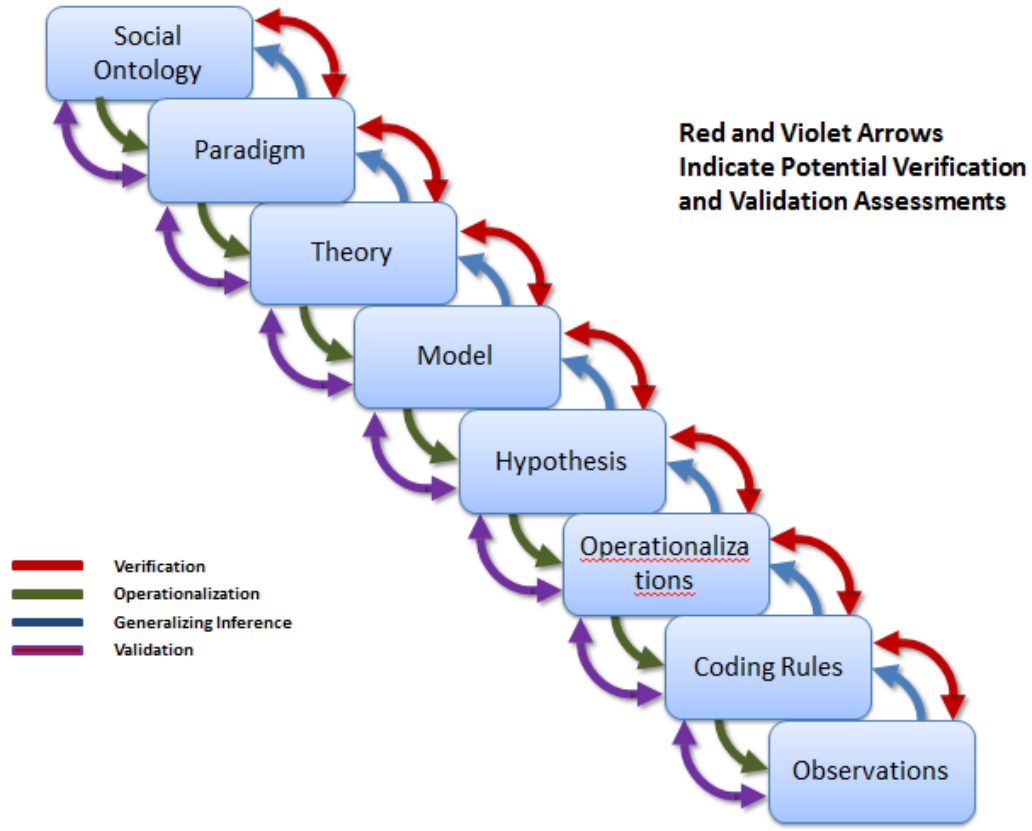
Operationalization and Substantive Inference as Generic Scientific Objectives

Within this framework verification and validation can be understood as assessments of up or down moves on this hierarchy from one level or "rung of the ladder" to another. In other words, one might identify the source of an error in a model's performance as an inconsistency between the "theory" that is, in fact, a valid theory, and a model based on the theory that is an

incorrect specification of it. This would be a "down verification" error. Alternatively, one might attribute an invalid hypothesis to corroborative inferences supporting the hypothesis that were based on misapplied statistical measures of significance. This would be an "up verification" error. An example of a "down validation" error would be the failure of observations in a particular domain to support a hypothesis derived correctly from a valid. An "up validation" error would be signaled by failure of a theory to operate successfully in a domain more general than those validly explained or accounted for by the individual models from which the theoretical claim was correctly inferred.

Figure 3 registers the location of these operations as analogous at every level of the epistemological hierarchy.

Figure 3



Verification and Validation of Operationalization (Down) and Generalizing Inference (Up)

Verification as Construct Validity

The categories and stipulated definitions of Verification and Validation registered in the DoD's authoritative 2006 document can be traced most directly to DoD discussions and decisions that took place in the late 1960s. These discussions were prompted, in part, by the influence but also the dramatic limitations of the kinds of systems analysis and operations research models deployed by Secretary of Defense McNamara. Inspired by this challenge, analysts sought ways to evaluate the 'credibility" of mathematical or computer simulation

models used to assist in training or problem solving tasks associated with combat or resource allocation. In these discussions "verification" referred to a test of the "internal consistency" of the model. "Validation" referred to a test examining the extent of agreement between model outputs and something external—the real world or the output of another, presumably validated, model.[16]

Discussions of how to assess the credibility of these models generated much intense and often confused discussions. Participants found it very difficult to agree on terminology, but more importantly, on the concepts. In particular it was difficult to agree on a "model" for verification and validation that was itself clear enough, consistent enough with available theories of scientific evaluation procedures, and corroborated through experience, that as a model of verification and validation itself, it could be considered verified and validated, or at least verifiable and validatable. At long last drastic measures were adopted. In her authoritative treatment Sanders noted that these concepts—verification and validation--and their conventional definitions were as much the product of coercion as analysis: "Ten or so senior practitioners [were] locked in a room for days until they could agree."[17]

Among the specific problems identified by analysts of these discussions was that a model based on the correct implementation of a blueprint for the model could well fail verification tests if the model development concept (blueprint) itself was not a correct expression of valid, higher level theory. In other words, a model might seem to be invalid, even though verified as a faithful operationalization of the design concept. This could occur either because the design concept was not a verified operationalization of the valid theory behind it, or because it was a verified operationalization of an invalid theory. If the former, verification of the model would conceal inherited verification errors in the production of the design concept. If the latter, verification of the model would conceal inherited validation errors. Either way, the flaws in the model could only be discoverable through validation tests. Davis described the practice of assessing the process of producing the design concept for a model as "structural validity" meaning that "the model has the appropriate entities…attributes and processes so that it corresponds to the real world (verisimilitude) at least as viewed at a particular level of resolution."[18]

What is important to note here is how validation and verification are inextricably bound up with one another, showing again that once the design concept for a model is problematized, a categorical distinction between validation and verification becomes extremely difficult. Twelve years later, Sargent's influential work on validation and verification reflected the same difficulty, if not impossibility, of adhering to a clear distinction between "building the thing right" and "building the right thing" once questions are posed about the integrity of the design concept for a model. Sargent offered two definitions of "conceptual model validation" in the same article. The first describes conceptual model validation as assessing the "validity" of governing theories and the "reasonableness" of the operationalization of those theories (which would mean a combination of the "right thing having been built" and the "thing having been built right"):

> Conceptual model validation is defined as determining that the theories and assumptions underlying the conceptual model are correct and that the model representation of the problem entity is "reasonable" for the intended purpose of the

model (Sargent, 2004).[19]

His second formulation, however, combines what DoD currently treats as "verification" and "validation" but in the reverse order:

> Conceptual model validation is defined as determining that the theories and assumptions underlying the conceptual model are consistent with those in the system theories and that the model representation of the system is "reasonable" for the intended purpose of the simulation model (Sargent, 2004).

Thus for Sargent, conceptual model validation asks whether the model was built right, as a specification of the governing theory; *and* whether the model works, in a "reasonable way" given its purpose, *i.e*. that the right thing was built. This usage, so contrary to what has become more or less standard, is less surprising than it might be if it is recalled that in the late 1960s the operations now normally referred to as "verification" and "validation" were in fact combined under the one concept of "verification."[20] Indeed the tendency of deep thinking about verification and validation to lead to the conclusion that these two types of evaluation cannot in fact be completely disentangled from one another, or categorically distinguished from one another, can be traced back earlier to Schlesinger, *et. al.* (1979).[21]

Terminological confusion also arises as a result of the rather specific association of the term "model validation" with substantiation that a computerized model within its domain of applicability possesses a satisfactory range of accuracy. It is instructive to note how this formulation subtly blends verification and validation, since evidence that a computer model was not satisfactorily accurate could lead to hypotheses both that the program had not been written correctly to specifications ("the thing had not been built right") or that the program chosen to simulate a particular part of the real world was not adequate to the task even if built to specifications ("the right thing had not been built"). Additionally, we should note that the validity of the specification of the model's "domain of applicability" is taken as a prior judgment that is separate from model validation itself. But what sort of operation is entailed in determining the "domain of applicability" of a model? If that specification is produced by inferring boundary conditions from the model design, then we would consider the validity of the specification of the "domain of applicability" to be a "verification" operation. But if we consider the correct specification of the domain of applicability to be an empirical question, it would be treatable as a matter of generalizability, *i.e.* of "validation."

This inquiry—into how questions about "verification" morph naturally into validity questions—highlights the unavoidable fact that building a wrong thing the right way is one way to build the wrong thing. We can thus understand the terminological confusion cited above, highlighted by Bankes (2011), as associated with a fundamental conceptual problem afflicting attempts to distinguish categorically between assessments of coherence and assessments of empirical correctness. One key to escaping from the terminological morass that has long afflicted discussions of validation and verification *and* to integrating work done on the problem across DoD, natural science, social science, and academic domains, is recognizing that verification is nothing more nor less than "construct validity." Originally known as "congruent validity," construct validity is the judgment that two expressions of an idea are isomorphic to one

another within acceptable limits.[22] To be sure that a plan or a design concept has been implemented according to specifications is to check that the object to be designed and the object actually designed correspond to one another along important metrics. If the design of a factory specifies an outer wall as 100 meters in length, the length of the wall as it is built should be measurable as 100 meters (not 101 meters or 100 yards). To be sure that a variable's effect is being assessed correctly means insuring that the operationalization of the variable can be treated as the variable itself without loss of meaning. To be sure that a theory built from inferences based on patterns in the performance of different models is warranted, as a general expression of patterns in the specific outputs of those models, is to confirm that logical missteps have not been made in the process of inductive inference. To be sure, for theories to avoid conceptual contradictions their key terms must be found within a paradigm whose ontological foundations permit the objects referred to in the theory and the statements being made about those objects. Each of these "construct validity" operations, and any other that one can imagine, are "verification" operations, checks to see that "the thing was made right."[23]

With this principle in mind an instructive convergence appears in prevailing approaches within DoD and academia to validation of social science models. In each setting there is a significant displacement of attention from validation in general, to construct validation (verification); from the kind of validation focused, per se, on the substantive accuracy or generalizability of claims to the kind of validation focused on the warrantability of claims. In the Zacharias, *et. al*. study that appears to have been the basis for the 2009 DoD BAA, one does find discussion of validation techniques for IOS models intended to evaluate the performance of models by comparing their outputs to the implicit models used by experts using qualitative methods, to the outputs of extant quantitative models, and to the conclusions of both computational and field data based models targeting the same domain questions.[24] On the other hand, this kind of comparison and "triangulation" depends on prior belief in the accuracy and effectiveness of the other implicit or explicit models employed by these other techniques. Indeed the Zacharias, *et. al.* study puts most emphasis and goes into most detail on "docking" techniques for IOS models. Despite the origins of this metaphor in the "docking" of space vehicles, the idea of docking two models does not refer to connecting one to the other. Rather docking refers to an act of translation, of cross-calibrating a second model using the algorithms of a first model. The objective is to observe whether the outcome of the cross-calibrated second model in response to an identical challenge or question produces the same outcome or response as the first. However, this procedure is a check of construct validity, not of model accuracy. It is analogous to asking whether driving directions translated from English to Chinese will bring the driver to the same location; not whether the directions will bring the driver to his actual destination.

Recalling the text of the BAA cited above, it appears that this emphasis in the Zacharias, *et. al.* study, along with whatever other studies contributed to its framing, led to a virtually complete substitution of "verification" (construct validity) for "validation" (tests of accuracy or generalizability to empirical reality). This substitution of assessments of "warrantability" for assessments of accuracy or generalizability can be roughly understood as a displacement of interest from what is commonly understood as "external validity" questions to "internal validity concerns."[25] It is interesting to note that the same pattern of displacement is exhibited among social scientists, and in particular political scientists.

Applying the Unified Epistemological Hierarchy Model of Validation and Verification: A Plausibility Probe

I turn now to brief consideration of the high profile and prolonged debate between "Quals" and "Quants" in political science to see how the critiques both sides offer of the work of their methodological rivals are best understood as judgments about "construct validity" (i.e. verification), rather than critiques of the substantive accuracy or generalizability of one another's models. An outsized proportion of all methodological disputes within political science over the last three decades has pertained to one of two related arguments—one over rational choice and its game-theoretic relatives and the other over statistical manipulations of aggregated data. The adversaries in these two debates share claims to the mantle of science. In other words, although they divide along the "cultural" lines identified by Snow, they each evaluate their own methods and those of their adversaries according to what they identify as "scientific" norms. Still, patterns in these debates reproduce in the political science world much of the rhetoric that dominated the struggle Snow observed between the two cultures of "literary intellectuals" and "scientists." Self-descriptions of one side of this debate include a number of adjectives used to modify the word "research"--qualitative, idiographic, small-n, discursive, or interpretivist. Labels for the other side include "quantitative, nomothetic, positivist, formal, aggregate data-driven, and technically sophisticated." The adjectives in these groups are neither individually accurate nor synonymous, but they are widely used.

The first group—let's call them the "Quals"—condemn the second group as social scientists with physics envy who generalize applications of econometric techniques and microeconomic theory to vastly inappropriate domains. The second group—let's call them "Quants"—condemn the first group as atheoretic, lacking rigor, conceptually imprecise, and unable to test their claims or demonstrate the cumulativeness of knowledge production. Notably, in contrast to Snow's depiction of the debate (between science and non-science), each side in the methodological wars of political science condemn the other for not living up to the norms of science. For Quals, Quants are "scientistic," not scientific. For Quants, Quals are ignorant or unappreciative of the rigorous requirements of the rules of evidence in science and insufficiently interested in the production, evaluation, and accumulation of general claims.[26]

The debate is conveniently and intelligently documented in two books. The first is *Designing Social Inquiry*, by Gary King, Robert Keohane, and Sidney Verba (hereafter KKV). Published in 1994, it is an enormously influential manifesto urging Quals to discipline their work by subjecting it to the requirements for inference that KKV identify with best practices among statisticians for generating and evaluating evidence based claims. Its three major pieces of advice for Quals are to 1) increase the amount of data they use, even if that means limiting their focus to topics which the aggregation of data in systematically analyzable ways; and 2) attach confidence intervals (degrees of uncertainty) to all claims they do advance; 3) avoid errors of inference made with qualitative evidence that can be identified as analogous to the inferential errors statistical theory is designed to prevent. The second book *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (second edition, 2010) was edited by Henry E. Brady and David Collier as a specific response to KKV (It even contains an essay by those authors

responding to early criticisms.).[27] Its three major points are that 1) standard quantitative practices contradict fundamental tenets of statistical theory; 2) high level ontological and theoretical commitments implicit in the Quant approach constrain its usefulness; and 3) advice offered by KKV ignores the scientific importance of concept formation and substantive theory development, substituting instead search for theoretically problematic but conveniently countable observations.

Considering each of these six critiques in turn, we can identify their locations on the epistemological hierarchy (high, in the "conceptual/theoretical" realm or low, in the implementation or "operational" realm), the character of the assessment they represent (validation or verification), and the direction of the assessment (up or down).

In their concluding chapter KKV identify finding "as many observable implications of your theory as possible" as the "primary way" to gain "leverage over research problems."[28] They repeatedly urge Quals to increase the amount of data available, even if that means privileging models and hypotheses that enable quantifiable inputs as evidence over models and hypotheses that do not. They ask "how many observations are enough?" The answer: "The more the better…," especially when testing the relatively imprecise models characteristically deployed by Quals.[29] In terms of the model of V and V presented here, this advice to Quals is based on identification of a propensity to commit up verification errors in operational ("low") rungs in the epistemological hierarchy where inferences from patterns in data are used to corroborate hypotheses.

Regardless of how much data Quals gather, however, KKV admonishes them to always be explicit about the degree of uncertainty attached to the claim being made. "Perhaps the single most serious problem with qualitative research in political science is the pervasive failure to provide reasonable estimates of the uncertainty of the investigator's inferences." (32) Only by verifying correct application of a variety of rules for making proper inferences from available data can investigators assess and report the amount of uncertainty attached to the claimed validity of the proposition under investigation. Among the most important errors made by Quals, KKV identify those pertaining to measurement validity (Are the indicators used sufficiently congruent to the concepts being operationalized?); multicolinearity (Are inferences linking an independent to a dependent variable weaker than they seem because of the contribution of other variables that vary long with the chosen independent variable?); endogeneity (Is the effect, as framed by the proposition, actually part of the cause?); and selection bias (Do sampling techniques produce unrepresentativeness in directions relevant to the proposition under investigation?). Located by KKV on low rungs of the epistemological hierarchy, these are errors said to be common in Qual work, work featuring non-credible generalizations from data to hypothesis, or hypotheses to models and theories. In the lingo of DoD such outcomes are failures to "build the thing right." In the language of the framework advanced here, these are "up verification" critiques, challenging the warrantability of claims, based on faulty inferential logics for constructing generalizations from concrete observations, not the accuracy of claims (assessment of which would entail "up validation").

In sum, *Designing Social Inquiry* offers itself as a manual for Quals that distills and explains the inferential errors statistical theory is designed to prevent statisticians from making.

In other words, it presents a model of political science research that is inferred from a theory of statistics. As I have shown, KKV use this model to engage in an "up verification" critique of Quals. Quals, on the other hand, focus an important part of their critique of KKV on "down verification" errors, faulting KKV, for errors of inference in their model's operationalization of statistical theory.

Brady, Collier and Seawright identify statistical theory as providing "an underpinning for the critiques focused on the application of KKV's ideas to qualitative research."[30] In a detailed critique of the enormous amount of Quant work done to identify relationships between economic development or performance and democracy, Jason Seawright shows that incorrect but prevalent misapplications of statistical theory rules by Quants have induced them to believe that multiple regression analysis could be used to identify valid explanations for patterns relevant to this relationship. Based on his "down verification" critique of moves by Quants from statistical theory to improperly generated, iteratively specified, and severely confounded regression models, Seawright explains the pattern of "remarkably inconsistent findings on this topic in the literature that employs conventional cross-national regression analysis."[31] Thad Dunning echoes Seawright, citing work by Chris Aachen and David Freedman as contributing to "[g]rowing recognition of the frequently severe problems with regression-based inference…which fall under the rubric…of mainstream quantitative methods." He cites Achens's well known "rule of three," based on the impossibility of controlling an analysis that includes as many interaction effects as would be associated with more than three variables. Dunning comments that conventional practice by Quants is a "far cry" from this advice, "in which the trend has been towards more complex statistical models in which the assumptions are difficult to explicate and defend…"[32] In other words, Dunning is arguing that aside from the nonconformance with some tenets of statistical theory, the increasing complexity of multiple regression models itself renders verification impossible. Describing himself as a "quantitative methodologist," Larry M. Bartels writes that he is "chagrined to notice how wobbly and incomplete are some of the inferential foundations that KKV claims are 'explicated and formalized clearly in discussions of quantitative research methods."[33]

A second type of Qual critique of the KKV approach can be understood as conceptual challenges located relatively high on the epistemological hierarchy. The statistical perspective operationalized by KKV treats the world as "probabilistic" rather than "deterministic."[34] The objection is that this problematic endorsement of one statistical theory over another automatically undervalues or even excludes crucial Qual methods for verifying their claims. As detailed by contributors to the Brady and Collier volume, process tracing is a central aspect of Qual methodology—the careful observation of chains of consequence in a particular case or episode. As Collier, Brady, and Seawright put it:

> qualitative researchers use causal-process observations, as we put it above, to slowly but surely rule out alternative explanations until they come to one that stands up to scrutiny. This is a style of causal inference focused on mechanisms and processes, rather than on covariation among variables.[35]

Causal-process observations (which Collier, Brady, and Seawright counterpose to KKV-advocated "data-set observations.") document the linkages necessary to warrant belief that a

correlation consistent with the hypothesis is in fact produced by the logic of the model that produced the hypothesis. In the example presented by David A. Freedman in the Brady and Collier volume, the theory of waterborne infectious diseases, inspired by the cholera model, produced a set of hypotheses that "plague, yellow fever, dysentery, typhoid fever, and malaria…were waterborne infectious diseases." However, "because guesses cannot be verified," careful process tracing of episodes of outbreaks of these diseases was necessary to show that some correlationally warranted inferences could not be interpreted as evidence of the validity of the theory. These studies show that in fact the theory of waterborne infection in fact did not apply to most of these diseases, only to typhoid and dysentery. 233 In his dissection and refutation of the argument that early media pronouncements of Gore as the winner of Florida in the Presidential election of 2000, Henry E. Brady shows explicitly how correlational arguments based on data-set observations lead to conclusions shown to be invalid by a focus on causal-process observations.[36] My point is that this kind of validation test, anchored in a deterministic, mechanism-focused rather than probabilistic approach to causality, is logically impossible or at least highly problematic from the probabilistic perspective.

Below ontology, but still in the higher "conceptual" runs of the epistemological ladder, the Quals level another critique at the Quant position. Although "frequentism" was the standard theory of statistics for decades, and is fully embraced by KKV, [37] Bayesian theory always has been and recently has become a potent rival. According to C,B and S, KKV rely overmuch on frequentist theories of probability as opposed to Bayesian approaches which do not require strong assumptions (often not met) about the shapes of population distributions. 166 This "down validation" critique—a challenge to the validity of the model used by KKV (insofar as it *is* correctly inferred from frequentist theory) as a basis for constructing their methodology model--also illustrates the third major challenge Quals level at Quants. KKV, and the aggregate "data crunchers" seen as the primary advocates and followers of their approach, ignore or drastically undervalue the contribution to science made by the development, refinement, and elaboration of concepts and theories—activities wholly apart from or only indirectly linked to measurement, observation, and data collection. Bartels's criticism of KKV highlights this displacement of interest from concept formation and theory building to counting. According to Bartels, KKV's nod to such activities is "patronizing," *viz.* "there must be room in science for "ideas regarding the generation of hypotheseses."[38] He goes on to cite KKV's advice to multiply "observations" at all costs, even if those are not "implications" of a substantively applicable theory, as reflecting a dangerous devaluation of the role of substantive theory.[39] The often cavalier attitude of the Quants toward theory and its role in producing models, and through them hypotheses to be tested, is well captured in the critique of the common practice among multiple regression modelers (alluded to above) of serial respecification of models, i.e. curve-fitting.

> A vast number of journal articles have sought to make causal inferences by estimating perhaps a dozen related (though quite typically under-theorized) model specifications, picking and choosing among these specifications, and offering an ad hoc interpretation of a few selected coefficients.[40]

Such criticisms of the KKV model reflect judgments that overemphasis on insuring against up-verification errors low in the epistemological hierarchy, results in or is associated with down-validation errors in the upper rungs, i.e. inadequate attention to the provenance and accuracy of

theories appealed to as guiding model specification and hypothesis generation.

Overall we can see that the Quant critiques of Qual work focus almost entirely on up-verification errors committed in the "operational" zone of the epistemological hierarchy. Qual critiques of dominant Quant methods focus heavily on down-verification errors in the operational zone of the hierarchy, mostly pertaining to what is commonly known as "measurement validity," while also highlighting down verification and validation errors located in the upper ranges of the hierarchy—the "conceptual" zone, where concepts and high level theories are serve as sources for models. In this context it is useful to recall the implicit judgments of the ONR BAA, namely, that social science models cannot be "validated" against historical other empirical evidence, but that verification of these models via establishing their links to higher level (reigning, if not validated) theories can help enhance their credibility. Essentially, the BAA and the Zacharias, *et. al.* report upon which it appears to have been based, can be understood as echoing the insistence of the social science "Quants" that operational level "verification" efforts, through careful application of the norms of statistical theory, are to be preferred for assessing the credibility of social science models over validity testing of actual causal claims. Corresponding operations suggested in the Zacharias report are a variety of docking, documentation, and triangulation techniques that conform to the BAA mandate for verification over validation. On the other hand, the emphasis of the BAA on down verification of models based on their relationship to higher level theories, echoes a key element in the Qual critique of Quant work as curve-fitting and as only loosely disciplined or informed by substantive theory.

One big difference between the DoD discourse on V and V, and the implicit debate over these operations we have analyzed among social science methodologists, is in the motivation for embracing verification (as construct validity). While DoD suggests verification (by whatever name) as a means to enhance the credibility of social science models because social science models and natural science models are so different that "traditional" validation is impossible; social scientists (both Quants and Quals) engage in verification exercises because they agree that science is a unified field methodologically and that its rules can and should be enforced in the social sciences as they are in the physical sciences.

Whether the framework presented here for mapping verification and validation requirements is the most effective way to organize these activities across scientific domains and across government and academically-based communities, the need for such a framework is highlighted by a culturally driven disconnect in terminology. While DoD insists in its protocols on "validation" and "verification," social science methodologists almost never use the term "verification." One argument presented here is that there are good conceptual and theoretical reasons for considering verification as a form of validation. On the other hand, "construct validity," which is what verification actually means for DoD, tends to get lost as a general kind of assessment, amid a welter of overlapping and underspecified types of "validity" that appear in the lexicon of social science methodologists.[41] The explanation for this difference across communities in their attachment to the words "verify" and "verification" likely lies in the influence of computer science and engineering in DoD vs. the pre-falsificationist notions of "verifiable" and "verified" familiar to most social science methodologists. In this regard, the argument advanced here is an effort to bridge a cultural divide, as well as build a coherent analytic position.

[1] Department of Defense. 2006. Key Concepts of VV&A. [online] Modeling & Simulation Coordination Office. Available at: < http://vva.msco.mil/Key/key-pr.pdf>fe
[2] ONR BAA Announcement Number 09-026, "Human Social Culture Behavior Modeling," Department of the Navy, Science and Technology, section 6c, p. 5. (emphasis added)
[3] C.P. Snow, *The Two Cultures and a Second Look* (Cambridge: CUP, 1964) pp. 69-70.
[4] Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1970) Second Edition, p. VIII.
[5] *Ibid.*, p. 49.
[6] *Ibid.*, p. 92.
[7] Imre Lakatos, "Falsification and the Methodology of Scientific Research Programmes," in *Criticism and the Growth of Knowledge*, Imre Lakatos and Alan Musgrave, eds. (Cambridge: Cambridge University Press, 1970) p. 176. In his note to this passage Lakatos does actually use the term "social scientists" but only to accuse them of charlatanism. "[O}ne wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phony corroborations and thereby a semblance of 'scientific progress' where, in fact, there is nothing but an increase in pseudo-intellectual garbage." P. 176n.
[8] See for example, Lakatos, *op. cit.*, p. 108,
[9] Alan Musgrave, "Method or Madness?" in R. S. Cohen *et. al.*, eds. *Essays in Memory of Imre Lakatos* (Dordrecht, Netherlands: D. Reidel, 1976). For details on this point see Ian S. Lustick, "Lijphart, Lakatos, and Consociationalism," *World Politics*, Vol. 50, no. 1 (October 1997) p. 89. Political scientists seeking to distinguish natural science from social science ontologies as a basis for distinguishing two different kinds of science have recapitulated the pattern identified here, of actual embrace of a single scientific space even as the opposite claim is being advanced. See especially Gabriel A. Almond and Stephen J. Genco, "Clouds, Clocks, and the Study of Politics," *World Politics*, Vol. 29, no. 4 pp. (July 1977) pp. 489-522.
[10] *Behavioral Modeling and Simulation: From Individuals to Societies*, Greg L. Zacharias, Jean MacMillan, and Susan B. Van Hemel, Eds. (National Research Council, 2008).
[11] *Ibid.*, pp. 316-17.
[12] *Ibid., p. 317.*
[13] *Ibid.*, pp. 159-60.
[14] *Ibid.*, p. 323.
[15] *Ibid.*, pp. 8, 253, 289n, and especially 345.
[16] Clayton Thomas, "Verification Revisited, *Military Modeling for Decision Making*, ed. Wayne P. Hughes, Jr. (Alexandria, VA: Military Operations Research Society, 1997) pp. 334, 337.
[17] P. A. Sanders, "Accreditation: An Ingredient for Decision Making Confidence," in *Military Modeling for Decision Making*, ed. Wayne P. Hughes, Jr. (Alexandria, VA: Military Operations Research Society, 1997) p. 352.
[18] P. K. Davis, "A Framework for Verification, Validation, and Accreditation," In *Simulation Validation Workshop Proceedings, SIMVAL II*, ed. A.E. Ritchie. Alexandria, Virginia: Institute for Defense Analyses, 1992) p. VI-5.
[19] Robert G. Sargent, R. 2004. "Validation and Verification of Simulation Models," in *Proceedings of the 2004 Winter Simulation Conference*. eds. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters (2004) p. 1.
[20] Thomas, *op. cit*., p. 349n.
[21] S. Schlesinger, *et al.* 1979. "Terminology for Model Credibility," *Simulation*, Vol. 32, no. 3 (1979) pp. 103-104. The same may be said for the hoary and closely related distinction that originated with Donald T. Campbell between "internal validity" and "external validity." Campbell himself eventually abandoned the distinction is unworkable for the purposes for which it had been adopted. For a close treatment of this issue see Maria Jimenez-Buedo, "Conceptual Tools for Assessing Experiments: Some Well-Entrenched Confusions Regarding the Internal/External Validity Distinction," *Journal of Economic Methodology*, Vol. 18, no. 3 (September 2011) pp. 271-282.
[22] Lee J. Cronbach and Paul E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin*, Vol. 52, no. 4 (1955) p. 281n.
[23] For an in depth discussion of "verification" as the form of validation known as "construct validity," see Ian S. Lustick and Matthew R. Tubin, "Verification as a Form of Validation: Deepening Theory to Broaden Application of DoD Protocols to the Social Sciences," Paper presented at the 2012 Applied Human Factors and Ergonomics Conference (San Francisco: July 22-25, 2012).
[24] Zacharias, *et. al.*, 322.
[25] But again see Jimenez-Buede, *op. cit.*, for significant difficulties with this distinction.

[26] Postures taken by rational choice formal modelers and statistically oriented modelers working with aggregate data sets are quite similar, but sufficiently distinct that for the purposes of this analysis I will be focusing only on the latter as "Quants."
[27] Henry E. Brady and David Collier (eds.) *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Lanham, Maryland: Rowman & Littlefield Publishers, 2010).
[28] *Designing Social Inquiry, op. cit.,* p. 208.
[29] *Designing Social Inquiry, op. cit.,* pp. 216-17.
[30] David Collier, Henry E. Brady, and Jason Seawright, "Critiques, Responses, and Trade-offs: Drawing Together the Debate," in *Rethinking Social Inquiry, op. cit.,* p. 142. See also Henry E. Brady, David Collier, and Jason Seawright, "Refocusing the Discussion of Methodology," in *Rethinking Social Inquiry, op. cit.*, p. 26; and David Collier, Jason Seawright, and Gerardo L. Munck, "The Quest for Standards," in *Rethinking Social Inquiry, op. cit.,* p. 39.
[31] Jason Seawright, "Regression-Based Inference: A Case Study in Failed Causal Assessment," in *Rethinking Social Inquiry, op. cit.,* p. 247.
[32] Thad Dunning, "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" in *Rethinking Social Inquiry, op. cit.,* p. 274. For a critique of KKV as offering "homilies" (proverbs) disguised as a model improperly inferred from statistical theory, see also Henry E. Brady, "Doing Good and Doing Better: How Far Does the Quantitative Template Get Us?" in *Rethinking Social Inquiry, op. cit.*, pp. 67-82.
[33] Larry M. Bartels, "Some Unfulfilled Promises of Quantitative Imperialism," in *Rethinking Social Inquiry, op. cit.,* p. 86.
[34] *Designing Social Inquiry, op. cit.,* p. 145. By deterministic they say they refer specifically to "models of necessary and/or sufficient causation."
[35] David Collier, Henry E. Brady, and Jason Seawright, "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology," in *Rethinking Social Inquiry, op. cit.,* p. 191.
[36] Henry E. Brady, "Data-Set Observations versus Causal-Process Observations: The 2000 U.S. Presidential Election," in *Rethinking Social Inquiry, op. cit.,* pp. 237-242.
[37] There is no entry for "Bayes" or Bayesianism" in the index of KKV's *Designing Social Inquiry*.
[38] Bartels, in *Rethinking Social Inquiry, op. cit.,* p. 86.
[39] *Ibid.,* pp. 87-88.
[40] David Collier, Henry E. Brady, and Jason Seawright, "Introduction to the Second Edition: A Sea Change in Political Methodology," in *Rethinking Social Inquiry, op. cit.,* p.,9. See James D. Fearon and David D. Laitin "Integrating Qualitative and Quantitative Methods," Oxford Handbook of Political Methodology, eds. Janet M. Box-Steffensmeier and David Collier (Oxford: Oxford University Press, 2008) for a Quant response to this critique, acknowledged as potent. See also Laitin's very positive assessment of the *Designing Social Inquiry* even as he acknowledged that "is at its weakest in analyzing the role of concept formation in political science." David D. Laitin, "Disciplining Political Science," *The American Political Science Review*, Vol. 89, No. 2 (June 1995), p. 456.

[41] Adcock and Collier counted "thirty-seven different adjectives that have been attached to the noun 'validity' by scholars wrestling with issues of conceptualization and measurement." Robert Adock and David Collier, "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research" in the *American Political Science Review*, Vol. 95, no. 3 (Sept. 2001) p. 530.