CHAPTER XX

Verification as a Form of Validation: Deepening Theory to Broaden Application of DOD Protocols to the Social Sciences

Ian S. Lustick, PhD and Matthew R. Tubin, PhD

Lustick Consulting
Narberth, PA, USA
ilustick@lustickconsulting.com

ABSTRACT

The original DOD objective which the process of Verification, Validation, and Accreditation exists to serve was to evaluate the credibility of models and simulations. The red-line between verification (following the blueprint correctly) and validity (accuracy for a particular domain) was convenient, but suppressed consideration of flaws in design concepts. Protocols for evaluating models in domains where theory or data is not unchallenged, e.g. social science, requires clarifying the meaning of verification as "construct validity" and developing a model of verification and validation which situates these operations across multiple levels of an epistemological hierarchy. Pathologies of either inference or generalizability can present themselves at any level and be inherited by others.

Keywords: verification, validation, accreditation, social sciences

1 Deep in the Big Muddy: DOD Stipulated Verification and Validation

In his presentation at the Department of Defense's (DOD) HSCB Focus 2011 conference in February 2011, a noted authority on social science modeling—Steven Bankes—expressed his frustration at the terminological confusion relating to "verification and validation" in the social sciences. His solution was to abandon efforts to define existing terms in search of a new approach. Bankes's recent expression of frustration echoes a longstanding judgment about the failure of the social sciences to get its methodological ducks in a row when thinking about how to evaluate models or instruments. A welter of overlapping and sometimes synonymous terms are used in this context—including internal validity, external validity, measurement validity, face validity, construct validity, reliability, precision, convergent validity, and context validity. Over fifty-five years ago Cronbach and Meehl (1955, p. 281) observed that "Writers on validity during the preceding decade had shown a great deal of dissatisfaction with conventional notions of validity, and introduced new terms and ideas, but the resulting aggregation of types of validity seems only to have stirred the muddy waters." In the following six decades, not much has changed. In one influential study published in the *American Political Science Review* in 2001, the authors supported their assertion of continuing terminological confusion in the study of validation in the social sciences by noting that they had counted "thirty-seven different adjectives that have been attached to the noun 'validity' by scholars wrestling with issues of conceptualization and measurement" (Adcock and Collier, 2001, p. 530).

However, if "validation" has attracted considerable attention by social scientists, albeit with different meanings, "verification" is a term hardly ever used in discussions of how to evaluate social science models, theories, or procedures (Brady and Collier, 2010). It is not that social scientists have been unaware of the distinction between "building the thing right" (verification) and "building the right thing" (validation), but their approach to these problems has tended to collapse them into "validation" as a general, if confusing problem. As we shall see, the intuition behind this integration is correct (i.e. valid), though the logic for it is seldom understood (i.e. not verified).

DOD (2006) has stipulated, one might say "hard-coded," this distinction between "validity" of a model ("has the right thing been built?") and the "verification" of a model ("has the thing been built right?"). DOD uses this distinction to insure that the products and systems it acquires not only are built to specifications (verification), but also that they perform to requirements (validation). Passing scores in each category are required for "accreditation," the final step in the "V, V, and A" process. This language is clear with regard to the task of vetting and evaluating models and systems based on well-tested and well-corroborated natural science—using theories in relevant domains that have achieved consensual status in

the relevant scientific communities. However, as DOD has become more involved in the acquisition and deployment of social science based systems, significant problems arise. One difficulty is that in most domains of the social sciences models are derived from theories that are not consensually accepted. The Office of Naval Research within DOD has recognized the distinctive challenges posed by the need for V, V, and A on social science models and simulations. In 2009 it issued a BAA for HSCB modeling that included a focus on maturing, hardening, and validating human, social, cultural, and behavior (HSCB) modeling related software for integration into existing programs of record architectures, or maturing software via open architectures to allow broad systems integration (Office of Naval Research, 2009).

This paper has been written by Lustick Consulting staff as part of its support for Lockheed Martin ATL's Model Evaluation, Selection and Application (MESA) team. MESA's goals include constructing guidelines or protocols for best practice validation and verification for social science models and simulations. Accomplishing these objectives has meant thinking abstractly and precisely about the "constructs" that the standard DOD meanings of validation and verification are meant to operationalize. In other words, before we can determine effective V and V procedures for social science models, we must do a "construct validity" assessment of the DOD's own model for V and V. To what extent are the definitions quoted above, regarding "building the thing right" and "building the right thing" logical, efficient, and precise renderings of the evaluation objectives sought by the Department? Consideration of the origins of this terminology suggests that this model—DOD's official model of V and V—cannot, in fact, pass a validation and verification test if the model's intended focus of application includes social science models.

1.1 Constructing Credibility for Verification and Validation

The general conceptual problems associated with the DOD's sharp and fundamental distinction between "verification" and validation" were suppressed, not solved. We mean this literally. In her authoritative treatment of "Accreditation," Sanders (1997, p. 352) noted that these concepts 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." As DOD's interest in HSCB models expands and intensifies, these problems, conveniently set aside, can no longer be ignored. For unlike the natural science domain, where engineers regularly treat the theoretical basis of their model designs as incontestable, very few social science theories can be treated in this way. The fact is that theoretical claims upon which social science models are built are *normally* contested and have powerful rivals within the relevant community of social science experts. When experts do not agree on what the "right thing" is, determining that what is built has been built "right" cannot be categorically separated from tests of whether the "right

thing" has been built. For example, imagine a theoretically problematic model that is judged to be incorrectly built, i.e. not verified. Yet, its claims about the world are supported effectively by evidence. In this case it would be incorrect to infer from its performance support for the validity of theories governing the model or even for the model itself, generalized to other domains. In other words, the apparently clear distinction between "validation" and "verification" is blurred, a blurring that is more noticeable when theoretical knowledge is not consensual, e.g. when it comes to the evaluation of social science models.

The difficulty of maintaining a sharp distinction between verification and validation may be more apparent when evaluating social science models, but since no form of scientific knowledge is completely stable or consensual, the situation is no different fundamentally in the natural sciences. Imagine a natural science based model that does not work in the field, i.e. a model that is invalidated. One could then check to see that it has been built to specifications, including specification of its intended domain of application. If that "verification" check is successful, i.e. it is established with confidence that the model is a correct and faithful instantiation of the blueprint or "construct" produced by the designer, it can be inferred that the theory upon which the blueprint or construct was based was itself flawed, i.e. the science used by the designer to make the blueprint for the model/system was itself not valid. In that case the model would have inherited the invalidity of the theory upon which it was based. In other words, a verification check serves as a crucial element in the larger validation exercise.

The line of thinking here can be conveniently illustrated by considering the world of computer programming and the regular task faced by programmers of evaluating new programs (alpha and beta versions) for bugs. It is as if when it comes to validating and verifying social science models, every model should be treated as an "alpha" or "beta" version. When evaluating such "preliminary" versions of a computer program, computer scientists are fully mindful that the bugs they discover in the process of verification of the program may in fact be evidence of invalidity based on inherited problems in a flawed operating system or in the family of programs from which the program under examination has been derived.

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 (Thomas, 1997, p. 334, 337).

Among problems identified was that a model based on the correct implementation of a model development blueprint could well fail verification tests

if the model development concept 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 (1992, p. VI-5) 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."

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, the now standard and rigid distinction between validation and verification breaks down. Twelve years later, Sargent's (2004) 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).

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" (Thomas, 1997, p. 349).

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). Model validation is usually defined to mean 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 construct (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.

We are now in a position to return to our main question. How can the general problem of seeking credibility for models be conceptualized so that protocols applicable for both natural science and social science models can be stabilized and made as uniform as possible?

Recalling that originally both "verification" and "validation" were considered to be paths toward establishing or evaluating the credibility of a model, we suggest considering validation to mean evaluating the credibility of a model for an intended use. From a technical point of view, "verification" is the process for evaluating a model's credibility commonly known as "internal validity," more properly as "construct validity"—assessing whether the relationships between constructs and their operationalizations are clear and logically warranted (Adcock and Collier, 2001, p. 537; Morton and Williams, 2006, p. 9). "Validation" is the process for evaluating a model's credibility commonly known as "external validation" or assessing the amount of empirical corroboration for the accuracy of the model and the work that it does with respect to the real world. In deference to the conventions that have been established, we advise continued usage of "verification" and validation," but without the conventional belief that the two heuristics are

categorically separable and measurable independent of one another in all circumstances.

We still will require evaluative protocols and operations to be conducted for assessing coherence and fidelity to design concept (construct validation, i.e. verification) as well as for assessing the availability of corroborative evidence that the model is accurate and does work in the real world (external validation, i.e. validation). But we will not expect that these questions have categorical and complete answers at any one level of analysis. Put another way, if it is determined that a model fails a validation test by not conforming to patterns in the real world, one cannot assume that this failure resulted from poor implementation of the design concept (verification). A validation error may result from flaws in the theories that were the basis for the design of the model, which in turn may have resulted from logical or inferential errors in the process of generating the theory from either deductive or inductive processes of theory creation.

Perhaps it is the weaker faith that social scientists have in their theories, compared to their colleagues in the natural sciences, that explains the naturalness with which social scientists standardly *do* what DOD scientists standardly *did* before the reification of the "verification" vs. "validation" dichotomy, i.e. treat both the tasks of assessing the analytic equipment and the output of that equipment as important for the overall purpose of validating the "credibility" of a model. Note, for example, how a standard methodology text in the social sciences conceives of validation: "Validity," write King, Keohane, and Verba (1994, p. 25), "refers to measuring what we think we are measuring." For example, if we want to measure something in yards, can we "verify" that we have a yardstick and not a meter stick? Only by understanding "verification" as a form of "validation" (namely construct validity in all its forms, from "measurement validity" to theoretical operationalization of a model), can we develop an evaluative protocol general enough to be applicable to both the natural sciences and the social sciences, or to settings wherein theories are assumed to be true and when they are not.

2 CONCLUSIONS

Our key proposal can be expressed as follows. Verification (construct validity) pertains to the coherence of relations between operationalizations and constructs. External validity (validation) pertains to the generalizability of the content of substantive claims about the world. These are, as noted, distinctive heuristics for increasing or decreasing our confidence in a model. At any one level of analysis (see below regarding levels of analysis arrayed in an epistemological hierarchy), they will imply distinctive tasks to be performed and questions to be answered by the evaluator. To be sure, if one assumes or knows that the model design is empirically correct and perfectly clear, then its external validity should turn *only* on verification of the model. Empirical corroboration for a model known to be a valid representation of a valid theory would be unnecessary. But to the extent that the construct (model design) being operationalized is treated as potentially invalid itself

(i.e. theoretically flawed) or to the extent that it is not clearly and coherently stated, evaluation of the model cast as an operationalization of that construct will require investigation on multiple analytic levels.

This requirement implies need for a framework for parsing distinctive but systematically related levels of analysis. For models can present themselves on a continuum from very abstract or very concrete. A model for skill or aptitude testing may specify techniques for gathering and processing information. A model for translating social theory into categories of social structure in different tribal areas will operate at a much more abstract level. For this reason, unless model designs are assumed to be valid, the operations required to evaluate model credibility cannot be confined to any one analytic level. Accordingly, 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 we have developed, as part of the Lockheed Martin ATL MESA team, a ladder of locations ranging from extremely abstract to extremely concrete. At any one level the higher rung in this ladder is the "construct" or "model design" and the lower adjacent rung is the "operationalization" of the construct, or the implemented "model."

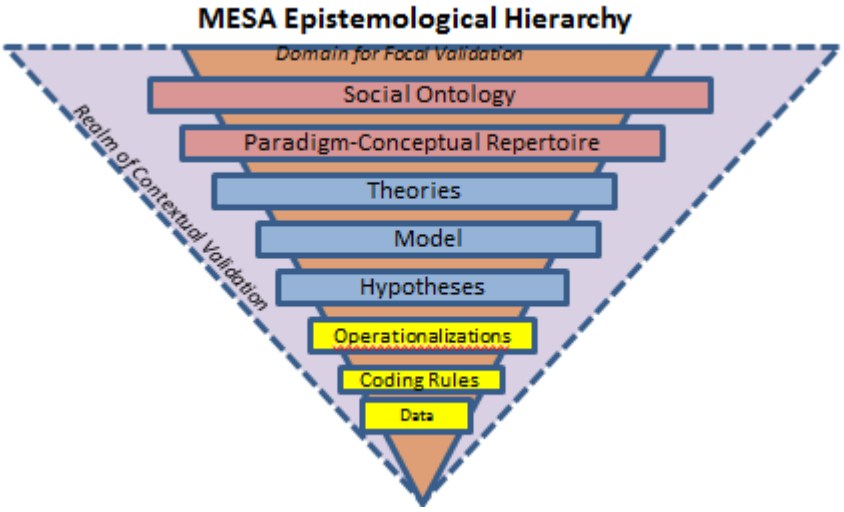

Figure 1 MESA Epistemological Hierarchy. For more information please see: A. Ruvinsky, J. Wedgwood and J. Welsh (2012).

The epistemological hierarchy, linking overall ontological or social metaphysics to the data processed as such by coding rules for generating or testing hypotheses, is displayed in Figure 1. The hierarchy, or ladder, of tasks that appears in the figure is simply a breakdown of the generic characteristics of the provenance or implications of any scientific claim. These are displayed within a focal range of immediately intended applicability and a wider range of speculative contextual applicability.

When we consider the validation and verification of a model we may in fact be asking about any rung on this ladder, with a paradigm (as a "model") built on the basis of higher level constructs (social metaphysics); a theory (as a model) produced from a higher level construct (paradigm); a model produced from a higher level construct (theory, i.e. design concept); an hypothesis produced from a higher level construct (model); operationalization produced from a higher level construct (hypothesis); coding rules produced in relation to a higher level construct (operationalization); and data produced from observations categorized according a higher level construct (coding rules). This account of the epistemological ladder is generalizable across all scientific domains. It has been presented from a top-down, "deductivist" perspective. But science includes just as prominently inductivist, bottom-up, processes. We may just as well be interested in validating and verifying the production of a higher level construct, such as a model, by combining and abstracting from an array of corroborated hypotheses.

Having established verification and validation as two approaches to establishing the credibility of a model, or identifying pathologies responsible for under-performance, we can design verification and validation protocols for any model, or any rung on the ladder. A big payoff for this approach is that the same protocols "valid" for one rung on the ladder will be valid for all—whether entailing "up-verification" and "up-validation" (questioning the persuasiveness and coherence of "theory-building" operations looking from one rung upward to an adjacent rung), or entailing "down verification" and "down-validation" (questioning the propriety and accuracy of "predictive" operations looking from one rung downward to an adjacent rung). By coherently organizing the evaluative tasks associated with establishing the credibility of a model we can both exploit the differences between verification and validation operations at any one level in the hierarchy, while appreciating the extent to which verification or validation weaknesses at a particular level may lead to the discovery of either verification or validation pathologies elsewhere.

ACKNOWLEDGMENTS

This work has been supported by a contract with Lockheed-Martin, ATL as part of its work for ONR under the MESA program (00014-10-C-0314). The views advanced in this paper, however, are solely those of the authors.

REFERENCES

Adcock, R. and D. Collier. 2002. Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95, 3: 529-546.

Bankes, S. 2011. Implementing Deep Validation. Presentation in *HSCB Focus 2011: Integrating Social Science Theory and Analytic Methods For Operational Use.*

Brady, H. and D. Collier eds. 2010. *Rethinking Social Inquiry: Diverse Tools, Shared Standards.* Lanham, Maryland: Rowman & Littlefield Publishers, Inc.

Cronbach, M. and P. Meehl. 1955. Construct Validity in Psychological Tests. *Psychological Bulletin* 52, 4: 281-302.

Department of Defense. 2006. Key Concepts of VV&A. [online] Modeling & Simulation Coordination Office. Available at: < http://vva.msco.mil/Key/key-pr.pdf> [Accessed: 13 February 2012].

Davis, P. K. 1992. A Framework for Verification, Validation, and Accreditation. In *Simulation Validation Workshop Proceedings, SIMVAL II*, ed. A.E. Ritchie. Alexandria, Virginia: Institute for Defense Analyses, pp. VII-VI24.

King, G., Keohane, R.O., and S. Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, N.J.: Princeton University Press.

Morton, R. and K. Williams. 2006. Experimentation in Political Science. [online] Available at: http://www.nyu.edu/gsas/dept/politics/faculty/morton/ExpChapHandbook5April06.pdf [Accessed: 14 February 2012].

Office of Naval Research. 2009. Human Social Culture Behavior Modeling. *BAA Announcement Number 09-026*. [online] Federal Business Opportunities. Available at: <https://www.fbo.gov/utils/view?id=ea459ff2dde33620486079 57ccc830a0> [Accessed: 14 February 2012].

Ruvinsky, A., J. Wedgwood and J. Welsh. 2012. Establishing bounds of responsible operational use of social science models via innovations in verification and validation. *4th International Conference on Applied Human Factors and Ergonomics*.

Sanders, P. A. 1997. Accreditation: An Ingredient for Decision Making Confidence. In *Military Modeling for Decision Making*, ed. W.P. Hughes, Jr. Alexandria, VA: Military Operations Research Society, pp. 351-59.

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, 1.

Schlesinger, S. et al. 1979. Terminology for Model Credibility. *Simulation* 32, 3: 103-104.

Thomas, C. 1997. Verification Revisited. In *Military Modeling for Decision Making*, ed. W.P. Hughes, Jr. Alexandria, VA: Military Operations Research Society, pp. 333-350.