AN EXPLANATION-BASED APPROACH TO VERIFICATION AND VALIDATION OF MODELS

D. E. STEVENSON

Abstract.

1. National Science Digital Library in a Nutshell

“The National Science Foundation’s (NSF) National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) program seeks to create, develop, and sustain a national digital library supporting science, mathematics, engineering, and technology (STEM) education at all levels — preK–12, undergraduate, graduate, and life-long learning” [?]. The vision for NSDL was explored in several workshops as a series of papers and monographs [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?] that (1) characterize digital library as meeting the needs of learners, (2) enabling parallel access, and (3) dynamic use of materials in digital format and reliable available of collections. The NSDL project began with a meeting at NSF in September, 2000 to introduce all digital library projects to one another with an eye towards identifying potential areas of collaboration and partnership.

1.1. Basic Concepts. The underlying principle of the NSDL is that the network should be the library. In any library the essential transaction is one in which a single user interacts with content. Digital environments offer much richer interaction in content, tools, and services, while services include the ability to search, refer, validate, integrate, create, customize, publish, share, notify, and collaborate, to name but a few. Learners — students, teachers, faculty, and those pursuing continuing education — “connect to learn” but must “learn to connect.”

By networking users and content with tools the digital library enables three chains of support.

(1) Users are supported by profiles enable the formation of learning communities, which could be short-lived or long-lived.
(2) Content is supported by metadata that enables the formation of customizable collections of educational objects and learning material.
(3) Tools are supported by common protocols or standards enable opportunity for the develop- ment of varied application services that enhance the value of the content for the learner.

The NSDL program is built on several working assumptions:

(1) There is a need for large content collections and the amount of learning resources on the Web increases daily.
(2) There are well-documented difficulties in verifying or validating material on the Web.
(3) The breadth and interconnectedness of the Web are simultaneously a great strength and shortcoming.
(4) Learning objects can be used, reused, repackaged, and repurposed. The ease of such recom- bination opens many issues intellectual property and digital rights management concerns.
(5) The situation encourages new models of content and new technologies.
(6) There is a need for an “organizational infrastructure” that facilitates connections between distributed users and distributed content.
An opportunity and a need to consider multiple business models of sustainability, particularly in the area of services offered by the digital library.

The success of NSDL will be measured by the extent to which the many projects can embrace this collective sense of identity and mission. From the integration requirement emerged the Core Integration Team with the responsibility for tracking development and is faced with sociological as well as technical and managerial tasks to accomplish. For example, issues of interoperability of components (collections and services) and coordination of protocols and standards across the user and provider base must be addressed.

1.2. Sustainability. An attractive scenario for the long-term management of the digital library is to place responsibility in the hands of a non-profit organization. There are approximately 4,000 institutions of higher education and 16,000 local K–12 school districts in the U. S. If an average contribution of $10,000/year would generate $200M annually in operating funds. But how creators will be compensated for their efforts?

- Authors of “fine-grained” content such as short applet tutorials or simulators could be offered recognition from peers that would be suitable and important “compensation”.
- Digital rights management technologies would allow the creators and purveyors of content to differentially price and/or repackage portions of “coarse-grained” material that has been disaggregated. (Some publishers have begun to offer custom runs of selected textbook chapters to professors.)
- Dyson [?] and more recently Shapiro and Varian [?] have observed that re-conceptualizing information as a service rather than a good offers the opportunity for new revenue streams that can be directed back towards content creators. This view suggests making services available individually or through affiliation with existing organizations such as professional societies. Sponsorship opportunities following the public broadcasting model are yet another possibility.

While it remains a central premise of the NSDL program that there should be long-term support and open access to core components of the digital library, the program does encourage proposals which explore alternative (and multiple) models of sustainability. Ultimately, open and restricted access must co-exist.

1.3. Changing roles and future directions. The evolution of the NSDL program and development of the national digital library for STEM education are likely to produce several trends. First, the traditional roles and relationships to one another of faculty and librarians are changing on campuses. Faculty increasingly will need to develop base level expertise in issues of library management such as cataloging, metadata tagging, and preservation. Library staff are likely to find that delivering externally-produced resources will be balanced by a need to deliver to the external world the products of faculty efforts on campus. There may be similar implications for teachers and librarians in the preK–12 sector. If these trends do emerge, then they have important implications for graduate programs that will provide the next generation of faculty and library and information specialists.

Libraries are more than just arbitrary artifacts: these artifacts have intellectual content. This fact is acknowledged by NSDL Pathways projects. Each project focuses on a specific area of knowledge; the CSERD Pathway is just one such pathway.

2. CSERD Pathway Project

The Shodor Education Foundation was chartered in May, 1994 with the mission of developing computational science educational (CSE) programs for K–12 through professional education. Shodor is recognized as a — if not the — leader in CSE education. One aspect of the Shodor web presence is the Computational Science Education Reference Desk (CSERD). Unlike Shodor’s own
in-house developed products, CSERD is a collection of contributed materials. This lack of control of quality is bothersome, being completely counter to the Shodor’s motto, “How do you know it’s right?” Correctness is crucial since the origin of the materials that need and ought to be in the NSDL may come from experts and scientists (not all high quality numerical models are appropriate “as is” for use in teaching), educators and teachers, or students themselves. Shodor’s experience has been that many computational educational materials (even “research quality” ones) we have examined are not valid from either a scientific or computational point of view. Some of the objects already exist in some part of NSDL, but without the educational support materials to make them effective as true learning objects the objects fail in their educational mission. For example, most objects do not have lesson plans, suggested sequences of activities, rubric for assessment, links to standards or other materials need to make the objects truly useful in the educational setting. On the other hand, some of the modeling lessons and computational materials already accessible in education as developed by students or teachers have educational content but have not been verified or validated.

The CSERD Pathway tack is to develop guidelines for correctness following the concepts of verification, validation, and accreditation (VV&A) that are used by the Department of Defense; Department of Energy; National Aeronautics and Space Administration; and National Institutes of Science and Technology among others. VV&A emerged from the operations research community in the 1970s and the issues of correctness are now more relevant than ever because simulations are more common and more weight put on their outputs in the decision-making process.

The ideas behind VV&A are not completely obvious; the terminology can be confusing and professional practice in DOD and DOE are such high consequence that the discussion of VV&A requires advanced statistics. For the CSERD project, the goal is develop practices that encourage the substantial review of models and simulations in light of the educational goals of such objects.

Part of the information that must be included in the NSDL is information about the evidence of effectiveness of the materials themselves, which can only come from the user community. We want to transform users into producers and active participants who add value to the very materials they find in the NSDL, while assisting in identifying or creating new materials for inclusion.

3. Verification, Validation, and Accreditation of Models and Simulations

The terms verification and validation causes confusion in the modeling and simulation world. When simulations were strictly disciplinarian, there was little need for distinction because everyone shared belief in the same models: i.e., engineering models were used by engineers who had a shared world view. The introduction of multidisciplinary models and simulations changes this shared viewpoint because we can no longer assume that everyone in the room understands or accepts discipline-specific models.

The Defense Modeling and Simulation Office (DMSO) web site, www.dmso.mil, is the central repository of information on V&V with The American Institute for Aeronautics and Astronautics (AIAA) and American Society of Mechanical Engineers (ASME) each having standards relating to disciplinary issues. In the CSERD Pathway project we use the DMSO definitions for verification and validation; accreditation in the Pathway is our own definition:

- Verification is the demonstration that the model is logically correct and follows from the physical and mathematical laws used. For a computer simulation, verification shows that the specifications are fulfilled. This is deductive correctness.
- Validation is the demonstration that the model correctly predicts the phenomena modeled. This is inductive correctness.
- Accreditation is the act of showing that the educational purpose of the model or simulation is achieved. This includes relating the object to the K–12 standards, identifying assessment criteria, and grade level where appropriate.
VV&A practitioners in government or industry undergo extensive training. Most validation work includes extensive and often esoteric statistical work. CSERD does not have the resources nor do the volunteers have the time to undergo such training nor engage in long term studies. Therefore, the CSERD group is developing online tools and a stable of examples to educate the reviewers on VV&A. The approach taken was that given by the author in [?], which is based on an extensive review of the current ideas in the philosophy of science and VV&A practice.

The CSERD Pathways project must rely on volunteer help. The organizational metaphor is that of a high-end scientific journal, with each area having “editors-in-chief” who will receive the materials and choose reviewers. The verification and validation reviewers can often be the same person; the accreditation reviewer must be an educator.

The goal in developing guidance is to provide simple, easy-to-follow concepts, the approach chosen being to find characteristics of verification and validation that would resonate with scientists and engineers when they review objects. The characteristics chosen were based on scientific explanations. Eight characteristics were chosen as proposed in [?], [?]. These were further divided into verification and validation characteristics.

3.1. Verification. The basic goal of verification is to show that a model or simulation is deductively correct by demonstration or evidence. The four characteristics we consider are (1) consistency, (2) justification, (3) lawfulness, and (4) reasoned.

3.1.1. Consistency. Consistency means “freedom from contradiction.” The foundations of science and mathematics are hypothetical if-then statement but scientists and mathematicians use the term in completely different ways. Scientists and engineers use if-then in its inductive (subjective) sense while mathematicians use it in the deductive (material) sense. For both groups, the if-then is false only when the hypothesis is true and the conclusion is false.

Consistency is the basic platform for testing: there a set of input-output values we take as correct, we run the simulation and check the output with the values. If the values are the same, then the test is passed; otherwise, it fails; This is the essence of a black-box test. First courses in biology, chemistry, and physics provide examples of inconsistencies if the models are pushed to their physical limits.

One example of inconsistency is the failure of an experiment to meet predictions (see [?]). Another example is the so-called sinusoidal pendulum: a pendulum whose time-variant behavior is the a simple sine curve. Such pendula do not easily validate because large displacements of a real pendulum is too large to demonstrate sinusoidal behavior.

3.1.2. Justification. According to the Routledge Dictionary of Philosophy, the term “justification” belongs to a set of terms that also includes “rational”, “reasonable” and “warranted” that do not have a commonly agreed definitions or relationships. For CSERD, we associate justification with a derivation or mathematical proof, or for software, justification is generally associated with testing. In general, anything termed evidence is justification.

Example: In the natural sciences, justification is closely related to the set of experimental data relating to the concept. This becomes more confused when statistics are required.

Example: Misapplication of principles are justification errors.

Example: Any mathematical error in a derivation is an example of unjustifiable results. Justification should be reasonably intuitive to people since the common meaning approximates the technical meaning.

3.1.3. Law-like. We think of law-like as indicating the scientific laws used in investigations. For example, the law of conservation of momentum.

Examples: An error here might be an inappropriate use of a physical law. The unwarranted generalization of an observation to a law is also an example.
Example: The homunculus was invented as a “person” who lived in the brain and was in charge of such neural functions as thinking and sexuality. Arguments invoking homunculi are always circular and hence, unlawlike.

3.1.4. Reasoned. One sense of reasoned is to consider the use of rules to infer a conclusion. A common example is with the syllogism:

\[
P \quad P \rightarrow Q
\]

\[
Q
\]

In words, from \(P\) and \(P \rightarrow Q\) we can infer \(Q\) and we have reason to believe \(Q\). But we need an external reason to believe \(P\). Therefore, in validation (second) sense, for someone to actually believe \(Q\), that someone must believe \(P\) as a reason to believe \(Q\).

Example: This might be a “context” error wherein we use an improper physical reason. We could probably get a goodly amount of examples by listening to first-year students in any science class: They may have facts but not necessarily do the reasons the facts cohere.

3.2. Validation. Whereas verification is concerned with rules, validation is concerned with the agreement between the model or simulation and the evidence produced from observations. Validation differs from verification in that validation is inductive. Kemeny [?] gives the basic ground rules:

1. There must be a general theory to provide context.
2. The general theory must be established.
3. There must be facts known independently of the facts to be explained.
4. The facts to be explained must be logical consequences of the above three factors.

For validation, we consider (1) coherence, (2) credibility, (3) organization, and (4) relevance.

3.2.1. Coherent. A basic tenet of science is that the body of knowledge should be coherent. Areas of knowledge are coherent just in the case that all the elements “fit” together. This is a central tenet of modern science.

Example: Researchers were very skeptical of the cold fusion results of Pons-Fleishman in 1989. Their results simply did not fit the established results.

Example: A form of incoherence is outliers. Is this a verification problem (such as instrument error) or a validation problem (an experiment in a new regime)?

Example: Michelson-Morley experiment partially removed the ætheir theory due to incoherent results [?]. The experiment required such massive changes to the æther theory that researchers no longer accepted the theory.

Counterexample: A long-standing caveat concerning evidence is, “The absence of evidence is not evidence of lack.” For example, experiments to determine the life-time of protons and other hadrons. Experiments to date put the minimum lifetime of hadrons greater than \(10^{32}\) years. This is not evidence that hadrons do not decay, only evidence that if they do decay then it must take at least \(10^{32}\) years.

3.2.2. Credible. Scientific observation has an epistemic side, with sufficient meaning and credibility to contribute to knowledge. This obligation to make observations relevant to theory suggests that there is an essential influence of background theories on the observations themselves. rewrite The theories we believe or wish to test tell us which observations to make and to describe the results of observations; that is, bringing out their informational content, will always be done in the language of the conceptual and theoretical system already in place. The influence of theory on observation is often seen as a threat to the objectivity of the process of testing and validation of theories. If theories are allowed to select their own evidence and then to give meaning and credibility to the observations, the testing process seems to be
unavoidably circular and self-serving. Observations can are used to disconfirm theories or at least undermine the theorist’s confidence.

**Note.** Credibility is a psychological, human thing. The results of the double-slit experiment just are; the issue is the interpretation and the credibility of the interpretation relies on human skill.

Examples: Cold fusion.

Example: Subject Matter Experts (SMEs). An example of SME intervention is the Space Shuttle Columbia tragedy. The model of debris forecast that the Columbia was wounded; the preponderance of the decision-making body chose to ignore the simulation, citing the simulation’s age and inappropriateness.

3.2.3. **Organized.** Since Ernest Nagel’s influential book [?], most discussions of unity of science have been cast in terms of reductions between concepts and between theories. Each science has a community agreed-upon standard arrangement of ideas: Observations lead to concepts that lead to constructs that lead to principles that lead to theories. Models and simulations should make use of this structure.

One way to display this information is shown in Figure 1 taken from [?]. The left column is the theory as taught in texts and the right hand column is the methodological skills used in the conduct. A gross generalization would put verification activities in the left column and validation activities linking left and right columns.

The natural types of problems that arise in organization issues are when the records and transformations are unusable. Knowledge claims are often due to the inappropriate claims from statistical practices.

3.2.4. **Relevant.** Relevance is the ability to retrieve material that satisfies the needs of the user. There are many examples of people focusing on irrelevant things when attempting to validate. A scientist might more aptly think of such-and-such is a causality relation: something may be observed but that observation is not connected to the problem.

Examples: Hidden variable formulations in physics.

3.3. **We Need To Keep VV&A In Perspective.** Workers without a historical perspective can misinterpret the VV&A process. Science has a self-correcting nature. If there was ever a true Hegelian process of thesis-antithesis-synthesis, then science is such a process. To put a human dimension on this, consider the comments of Dr. Virginia Trimble, Professor of Astronomy and History of Science, at the University of California, Irvine. She made this comment on The Science Show on the Australian Broadcasting Corporation’s The Science Show on May 20, 2000.

“I think one should start with the good [scientific “mistakes”] because many people don’t even realise that being wrong is not necessarily entirely bad and the classic example within astronomy is the steady state cosmology put forward by Bondi, Gold and Hoyle in 1948 ... [It] was already clear within about five years even to them ... that it was wrong but in that five or six or seven years the idea prompted an enormous amount of observing and an enormous amount of careful calculation that might very well not otherwise have been done ... [The] fact of having two competing hypotheses, the steady state and the evolutionary or “big bang”, as Sir Fred Hoyle termed it, into a science. That’s one definition of science. You have an idea that you can go out and test and have some hope of disproving or falsifying and by having two competing hypotheses cosmology and some deep sense became part of mainstream science.”

The point of her comment is that VV&A is an integral part of science and engineering; the current emphasis in VV&A makes these requirements more visible. Mathematics, too, has its twists and turns. Fermat’s last theorem is a good example. The story started in the 17th Century and finally ended in the 1995 when Andrew Wiles and R. Taylor completed Wiles 1993 proof.
Declarative

Focus Questions: Questions that serve to focus the inquiry about events and/or objects to be studied

World View: The general belief and knowledge system motivating and guiding the inquiry

Philosophy/Epistemology: The beliefs about nature of knowledge and knowing guiding the inquiry

Theory: the general principles guiding the inquiry that explain why events or objects exhibit what is observed

Principles: Statements of relationships between concepts that explain how events or objects can be expected to appear or have

Constructs: Ideas showing specific relationships between concepts, without direct origin in events or objects. The relationships are also called propositions.

Concepts: Perceived regularity in events or objects (or records thereof) designated by a label.

Events and/or Objects: descriptions of events or objects to be studied in order to answer the focus question.

Procedural

Value Claims: Statements based on knowledge claims that declare the worth or value of the inquiry

Knowledge Claims: Statements that answer the focus question(s) and are reasonable interpretations of the records and transformed records (or data) obtained

Transformations: Tables, graphs, concept maps, etc organizing records

Records: The observations made and recoded from events/objects studied.

To return to the historical nature of verification and validation, here is some “non-science” that was rejected by one or the other method but that led researchers to better science.

- Ptolomaic universe.
- Transmutation of lead to gold.
- Perpetual motion.
- Cold fusion.
- The æther theory of waves.
- Geocentrism.
- Malthusian population.
Maxwell’s Demon

The point, of course, is that VV&A is not about tearing down, but about building up knowledge.

4. Online Review Project

The CSERD Pathway has developed an online feedback mechanism that is available to all CSERD users. Any user can choose to add a review using a page similar to that shown in Figure 2. Reviews can be guided (Figure 2) or unguided (Figure 3). At the end, there is a review and approval (Figure 3).

5. VV&A in Action

A complete version of this example is available online based on the Force Concepts Inventory (FCI)[?]. Misconceptions have been shown to impede learning; by specifically attending to V&V issues, we can learn the depth of the misconception and attack the root cause.

5.1. A Two Falling Balls Problem. From the Force Concept Inventory[?].

Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the top of a two story building at the same instant of time. The time it takes the balls to reach the round below will be:

(A) About half as long for the heavier ball.
(B) About half as long for the lighter ball.
(C) About the same time for both balls.
(D) Considerably less for the heavier ball, but not necessarily half as long.
(E) Considerably less for the heavier ball, but not necessarily half as long.

1www.cs.clemson.edu/~steve/Papers/ICCSE
Figure 3. Guided Review Page

Misconception tested is G3: Heavier objects fall faster for 1A and 1D. (B) and (E) are logical restatements of (A) and (D), respectively.

5.2. **Model.** A model would have balls $B_1$ and $B_2$ with masses $\text{mass}(B_1)$ and $\text{mass}(B_2) = 2\text{mass}(B_1)$. Since the question put data in terms of weight, the student must have a model $\text{weight}(B_1) = g \times \text{mass}(B_1)$ and $\text{weight}(B_2) = g \times \text{mass}(B_2)$. The questions before the student is (1) what is the correct model for the time to fall from a height of $h$ (2) verify that formula and (3) validate the formula. Validation is tantamount to refuting the misconception.

The validation is the recognition that

$$F = ma = mg$$

and the masses $m$ cancels out, leaving

$$\frac{dv}{dt} = g$$

$$\frac{ds}{dt} = v.$$
5.3. Verification.

<table>
<thead>
<tr>
<th>Consistency:</th>
<th>The model is a standardly derived model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justification:</td>
<td>The model and derived equations are correct based on standard techniques</td>
</tr>
<tr>
<td>Law-Like:</td>
<td>The model was derived using the laws of motion and the kinematic principles in common use.</td>
</tr>
<tr>
<td>Reasoned:</td>
<td>The derivation uses accepted rules of mathematics.</td>
</tr>
</tbody>
</table>

Physics dictates that position \( s \), velocity \( v \), and acceleration \( a \) as shown in Eqn ?? and that system is solved as

\[
(2) \quad s = gt^2 + v_0t.
\]

Our initial conditions are \( g = 9.00665\, \text{m/sec}^2 \) (from the NIST website) and \( v_0 = 0 \), leaving

\[
(3) \quad s = 9.80665t^2
\]
5.4. **Validation.**

<table>
<thead>
<tr>
<th>Coherent:</th>
<th>The time it takes to fall in experiments does not differ significantly from times predicted by the model.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credible:</td>
<td>The model generates the correct time for the balls to fall given reasonable initial values.</td>
</tr>
<tr>
<td>Organized:</td>
<td>The model does not generate new theories and follows current versions of Newtonian gravitation.</td>
</tr>
<tr>
<td>Relevant:</td>
<td>The models answers the focus question concerning the time taken by two balls of differing weights. This may be seen by the graph of the position (model) or the table of experiments.</td>
</tr>
</tbody>
</table>

The major difference in *verification* and *validation* is that validation must square with reality. For this example, reality is dropping two such balls as $B_1$ and $B_2$ from a measured height, say 100 meters. This brings up proper statistical techniques and the proper formulation of the statistical model. We are hypothesizing that *(Hyp 1)* the time for $B_1$ and $B_2$ to fall are identical and that *(Hyp 2)* they both fall in the time predicted by the model Eqn ??.

These experiments can be done in any physics lab, at any educational level and with predictable results. Pedagogically, the dropping of the balls and the measurement of time should solve the misconception.
5.5. **Final Comments.** While the misconceptions problems provide good exercises in VV&A procedures, students holding a misconception may not be convinced one iota. To clear this misconception, it seems that a faster way to clear the misconception is to take two balls, weigh them, drop them, and hear for themselves. When students do that experiment, they should then be required to ask “Why?” and then provide the evidence.

315 McAdams Hall, Department of Computer Science, Clemson University, PO Box 341906, Clemson, SC 29634-1906

E-mail address: steve@cs.clemson.edu