Evaluation (Not Validation) of Quantitative Models

Naomi Oreskes*

Gallatin School of Individualized Study, New York University, New York, New York

The present regulatory climate has led to increasing demands for scientists to attest to the predictive reliability of numerical simulation models used to help set public policy, a process frequently referred to as model validation. But while model validation may reveal useful information, this paper argues that it is not possible to demonstrate the predictive reliability of any model of a complex natural system in advance of its actual use. All models embed uncertainties, and these uncertainties can and frequently do undermine predictive reliability. In the case of lead in the environment, we may categorize model uncertainties as theoretical, empirical, parametrical, and temporal. Theoretical uncertainties are aspects of the system that are not fully understood, such as the biokinetic pathways of lead metabolism. Empirical uncertainties are aspects of the system that are difficult (or impossible) to measure, such as actual lead ingestion by an individual child. Parametrical uncertainties arise when complexities in the system are simplified to provide manageable model input, such as representing longitudinal lead exposure by cross-sectional measurements. Temporal uncertainties arise from the assumption that systems are stable in time. A model may also be conceptually flawed. The Ptolemaic system of astronomy is a historical example of a model that was empirically adequate but based on a wrong conceptualization. Yet had it been computerized—and had the word then existed—its users would have had every right to call it validated. Thus, rather than talking about strategies for validation, we should be talking about means of evaluation. That is not to say that language alone will solve our problems or that the problems of model evaluation are primarily linguistic. The uncertainties inherent in large, complex models will not go away simply because we change the way we talk about them. But this is precisely the point: calling a model validated does not make it valid. Modelers and policymakers must continue to work toward finding effective ways to evaluate and judge the quality of their models, and to develop appropriate terminology to communicate these judgments to the public whose health and safety may be at stake. — *Environ Health Perspect* 106(Suppl 6):1453–1460 (1998). *http://ehpnet1.niehs.nih.gov/docs/1998/Suppl-6/1453-1460oreskes/abstract.html*

Key words: model evaluation, model validation, quantitative models

> Long experience has taught me that with regard to intellectual matters, this is the status of mankind: the less people know and understand about such matters, the more positively they attempt to reason about them.
>
> —Galileo

About lead in the environment, this much is certain: lead is bad. Human ingestion of lead is associated with a number of clinically well-documented afflictions, not least of which is the retardation of brain development in infants and children. Thus in the 1970s, the U.S. government began to take steps to decrease human exposure to ambient lead, most significantly by banning the use of lead additives in gasoline (*1–4*). Similar actions have been taken in other countries (*5*). Scientists working on the problem of assessing and regulating lead in the environment thus enjoy the benefit of widespread agreement about the basic harmfulness of the substance being regulated. (This is not to say that the consensus was not hardwon: In the 1920s and 1930s, most health professionals opposed banning lead in gasoline [*1,2*]).

The political and scientific consensus on the harmfulness of lead stands in contrast to other recent debates in environmental health and safety—nuclear power, polyvinyl chloride, radon gas, to name a few—in which there have been heated and even bitter disagreements among government agencies, industrial organizations, labor unions, and citizens' groups as to the significance of the purported harms (*6*). In these cases, debates have arisen in part because of the difficulty of documenting exposure levels (thus proving harm) in nonoccupational settings. Such settings typically involve low-level exposures whose clinical effects may be difficult to discern and characteristically emerge only after considerable time. In addition, the harmful materials may not themselves reside in the body and therefore cannot be directly measured. Under such circumstances, scientific uncertainty is inevitable. Low-level radiation is a case in point. Because radiation does not reside in the bloodstream, it is difficult to document exposures in uncontrolled settings, and impossible to prove that low-level exposure caused a particular affliction in a particular individual. Such proofs must rely on statistical regularities in longitudinal studies of populations. In contrast, it is relatively easy to document who has been affected by lead: blood lead levels are measurable and the clinical effects of toxicity are readily discernible (*7–11*). In principle, therefore, it should be a comparatively straightforward task to set legal limits for lead in the environment.

In practice, however, the problem of setting regulatory standards for lead has been complicated by the growing recognition that very low levels of lead exposure may not be safe as previously assumed (*2,3,12–14*). The problem of lead in the environment thus increasingly resembles other environmental health debates: the effects of low-level exposure—diminished school performance, attention disorders—may not be readily discernible and are difficult to diagnose. Even if accurately diagnosed, there is currently no safe medical treatment for low-level lead toxicity, and its most worrisome effects are irreversible. By

This paper is based on a presentation at the Workshop on Model Validation Concepts and Their Application to Lead Models held 21–23 October 1996 in Chapel Hill, North Carolina. Manuscript received at *EHP* 16 January 1998; accepted 13 May 1998.

I am grateful to L. Small for help interpreting the Ohio vs. EPA decision, and to R. Elias and A. Marcus for inviting my contribution to this volume.

*Present affiliation: History Department and Program in Science Studies, University of California-San Diego, La Jolla, CA.

Address correspondence to N. Oreskes, History Dept. 0104, U.C. San Diego, 9500 Gilman Dr., La Jolla, CA 92093-4695. Telephone: (619) 534-4695. Fax: (619) 534-7283. E-mail: noreskes@ucsd.edu

Abbreviations used: PVC, polyvinyl chloride; IEUBK model, integrated exposure uptake biokinetic model; U.S. EPA, U.S. Environmental Protection Agency.

the time a child is diagnosed with lead poisoning, exposure has occurred and damage has been done. Thus there is a compelling need to understand the effects of low-level lead exposure in order to prevent lead poisoning. Toward this end, scientists have turned to numerical simulation models.

Scientists at the U.S. Environmental Protection Agency (U.S. EPA) have been charged with the task of determining the relationship between environmental lead exposure and adverse health effects, with the goal of setting appropriate regulatory standards for lead in air, soil, and water in the United States. To address the numerous variables involved, the U.S. EPA has developed the integrated exposure uptake biokinetic (IEUBK) model, a software package consisting of several, linked computer programs that relate environmental lead exposure to blood lead levels in children (*15–18*). Model input consists of data on environmental lead exposures estimated by cross-sectional measurement of lead in air, soil, and water in children's homes. The data are fed into a biokinetic model that simulates the metabolism of lead in the children's bodies, and from this, estimates likely blood lead levels. In principle, the IEUBK model should be a powerful tool to help set nationwide regulatory standards, to identify communities in which current ambient lead levels are cause for concern, and to assess the likely impact of possible remedial actions in particular situations. In short, the goal of the IEUBK model is to prevent lead poisoning among American children, a goal that no right-minded person would dispute. But how do we know if the model is a good one? The demands of good science and the demands of democracy require evidence that the model is reliable (*19–21*).

Much of this demand has been expressed in terms of the need for model validation. As computer models are being used increasingly by federal, state, and local governments as a basis for policy decisions, there has been a concomitant demand for scientific agencies to attest to the legitimacy and reliability of these models, and to ensure that claims made on behalf of models are defensible. The safe level of lead exposure is a scientific question, but it comes to the fore within a social and political context. It was in this context that the U.S. EPA National Center for Environmental Assessment organized the October 1996 workshop titled "Lead Model Validation" to explore possible responses to the demand for evidence of the reliability of the IEUBK model. Scientists involved in the construction and use of the IEUBK model wanted to discuss what it means—or should mean—to call their model valid or to speak of its valid application (*22*). The title of the workshop presupposed both the necessity and the possibility of validating the IEUBK model, but organizers were also concerned with the question of whether one can validate a numerical simulation model at all, i.e, whether one can demonstrate that a model is reliable in advance of its use. Also at stake was the question of how the language of validation, that is, how we talk about what we do, affects both the process itself and our perception of it.

The purpose of this paper, emerging from that workshop, is to review the problem of uncertainty in the information obtained from complex models of natural systems in the context of the regulatory environment. This paper does not seek to offer specific recommendations on how to develop quantitative measures of uncertainty in any particular model. Such recommendations are best left to modelers themselves, and several recent papers offer such recommendations (*23–28*). Moreover, the notion of uncertainty quantification itself requires qualification. There are many sources of uncertainty in numerical models. Commonly, only a few are easily quantified, many or most are quantified only with difficulty, and several may be not be quantifiable at all. If a model is conceptually flawed, quantification of input uncertainty will not make the model reliable. On the contrary, quantification may surround such a model with an aura of credibility that it does not deserve. Yet the demand for credibility is real enough. The current regulatory climate has led to a situation in which scientists frequently feel pressed to argue the strength of their models, often beyond the degree to which they feel entirely comfortable. It is one thing to ask that scientists discuss the pros and cons of a model but quite another to demand that they declare the model valid. Apart from the internal demands of the scientific community, the push for model validation is a response to the political exigencies of our times. How should scientists, in the capacity of scientists, respond?

Working from a False Pretense: The Notion of a Validatable Model

In recent years, scientists in various disciplines have developed the notion of model validation to refer to the process by which scientists attempt to demonstrate the reliability of a computer model. Hodges and Dewar (*29*), in a report for the RAND Corporation on computer models used by the military to evaluate the efficacy of weapons systems in battlefield scenarios, make the distinction between two kinds of models: those that can be validated and those that cannot. To be validatable, in their words, the situation being modeled must satisfy four criteria: *a*) it must be observable and measurable; *b*) it must exhibit constancy of structure in time; *c*) it must exhibit constancy across variations in conditions not specified in model; and *d*) it must permit the collection of ample data.

Models in social and policy sciences generally fail to satisfy these criteria and therefore cannot be validated; that is, their reliability as a basis for prediction cannot be demonstrated. Because the systems are incompletely known and may change with time, a model that works well under one set of circumstances may fail under a different set of circumstances (*29*). In essence, such models are trying to "predict the unpredictable" (*30*). Bankes (*30*), also writing for RAND, concludes that the use of computer models for prediction in policy analysis is not only generally misleading but potentially dangerous, and in the case of battlefield scenarios, literally so. When used for prediction, these models provide only the illusion of certainty. At best, the result is a false sense of security, at worse, a dangerous hubris. Bankes advises that policy models should be used primarily in an explanatory mode, to explore the range and possible consequences of policy options, including worst-case scenarios. He notes that this normally requires the development of multiple models. Models sometimes produce results that surprise their creators, and in doing so elucidate unknown implications of known information and overt implications of covert assumptions. Nonpredictive models can be informative but only as long as they are used in question-driven rather than answer-seeking frameworks (*30,31*).

The RAND authors restrict their arguments to models in the policy domain and suggest that their caveats do not apply to the hard sciences in which model predictions can be experimentally verified. But this is an arguable point; many of the difficulties encountered in the social world also apply in the physical world. Oreskes et al. and Oreskes (*32,33*), in a discussion of computer models in the earth sciences, note that the criteria outlined above—measurability, accessibility, and temporal and spatial invariance—are precisely those

features typically lacking in the natural systems that scientists are increasingly exploring with computer models. The reason is evident: If a physical situation fully satisfied these criteria, there would be little need for a numerical simulation. It could be described, in most cases, with a small number of deterministic equations. Computer models are needed and have become increasingly common in the natural sciences precisely because scientists are grappling with complex systems involving multiple interacting variables that are difficult to access, hard to measure, and may change in space or time. Furthermore, the interrelationships between these variables may be indeterminate or at least not yet determined.

There are, of course, computer models that predict singular deterministic events in the natural world. Celestial mechanics provides an example: computer models are commonly used to predict the positions of celestial bodies. As the recent collision of Comet Shoemaker-Levy with Jupiter shows, models in this field are very successful. The location and timing of this collision was predicted to a high degree of accuracy more than a year in advance. One might thus claim that such models can be validated by reference to actual events—and have been. But models in celestial mechanics represent relatively simple physical systems in which the operative forces can be described by a small number of deterministic equations, and in which the variables (e.g., the mass of Jupiter) are measurable constants. Indeed, they are the exception that proves the rule because people have been predicting the positions of the celestial objects for millennia, long before the advent of digital computers. Computer models in celestial mechanics are a matter of convenience, not necessity.

Most models in the natural sciences are different. They involve data that are indeed variable and difficult to measure. Consider lead in the environment. Lead exposure and uptake may depend on lead concentrations in soil, water, air, and household dust; the size and quantity of lead paint chips in a household; the amount of soil or number of paint chips that a child eats; the amount of time a child spends outdoors; whether she washes her hands before eating and, if so, for how long she scrubs; and so on. Each of these variables is difficult to quantify. Indeed, if one could quantify by monitoring the amount of paint a child ate, one would be morally compelled to intervene to prevent further ingestion. (In practice, measurement of household dust is used as a surrogate for ingestion level, but children in the same household will have different levels of ingestion due to different patterns of behavior.) The input variables may also change with time and with the seasons, e.g., if a child spends more or less time out of doors; as the child grows up and changes his habits or begins to attend school; or unpredictably, if the family moves or has a change in its economic or childcare situation. Short-duration sampling of lead levels in a child's environment provides only an estimate of actual lead exposure, and this, in turn, delimits only the range of possibilities for actual lead uptake. Furthermore, the meaning of these variables may not be invariant. There is some evidence that the same exposure levels may produce different effects in different people, perhaps because of inborn or developmental contrasts in susceptibilities, nutrition, or synergistic effects with other elements in the environment (*34*).

We model systems like these precisely because of their complex nature, as a means for grappling with complex variables, and toward the important social goal of preventing future cases of lead toxicity. But in the process of constructing the model, we embed uncertainty, and, as the examples given above indicate, only some of this uncertainty can be estimated, much less directly measured. The issue of inborn susceptibility differentials, for example, is very poorly understood. Future research may lead to a better understanding of why different individuals react differently to the same exposure, but for now, uncertainty remains.

How does this embedded uncertainty affect the predictive reliability of the model? That is a question that cannot be established a priori. It can be established only through the actual use of the model. And this is why models of complex systems, whether in the social or the physical and biologic sciences, cannot be validated in the sense that the RAND authors imply. There is no way to demonstrate the predictive reliability of such models. To imply otherwise by using the language of validation is misleading. But if we cannot demonstrate the predictive reliability of the model in advance, then how should we be evaluating the merits and demerits of a complex numerical simulation model? One step in the process may be to realize that prediction is not as important as it is often thought to be, for predictive power is itself a fallible judge of scientific knowledge.

Limits of Prediction

The RAND authors cited above assume that models in the natural sciences can be validated because their predictions may be tested by observation in the natural world. In making this claim, they are implicitly invoking the hypothetico–deductive model of science, namely, that scientific theories can be thought of as statements that entail logically necessary deductive consequences: predictions. If the predictions of a theory come true, then we have warrant for faith in that theory. But this focus on prediction may be misplaced.

A fundamental problem with the hypothetico–deductive model, as many philosophers have realized, is that it assumes closed systems. A statement of the form p entails q works if and only if the statement describes a closed system. But a closed system is a philosophical ideal, not a natural kind. Real-life systems are never closed, and experimental tests inevitably embed hidden assumptions (*32,35,36*). Because these embedded assumptions may be faulty, a true theory may fail its experimental test. A famous example of this is found in the history of astronomy. Scientists in the 16th century suggested that if the earth orbited the sun, as Copernicus proposed, then the angular position of a given star would change during the course of the year as the earth moved through its orbit. But when astronomers looked for this stellar parallax, they found none—and they rejected the Copernican theory (*37,38*). Implicitly, they were assuming that the earth's orbit was large relative to the distance of the stars and that their telescopes were powerful enough to detect the changes that occurred. Both these assumptions turned out to be very wrong!

In the case of Copernican astronomy, scientists rejected a theory that turned out to be true, but what about the reverse? Have scientists ever accepted a theory on the basis of successful predictions but later discovered that the theory was false? To be sure. The alternative to Copernican theory—the Ptolemaic system—was confirmed by reams of observational evidence and scores of successful predictions of planetary events (*37*). Scientists in the 16th century had grounds for accepting the Ptolemaic system. Had it been computerized, its makers would have had every reason to call it validated (assuming that word had then existed). Yet, as we all know, the Ptolemaic system was fundamentally wrong. It was wrong not because it failed

its predictive tests but because the basic conceptualization of the universe that supported it was faulty.

In light of historical examples like this one, the philosopher of science Karl Popper famously argued that no scientific theory or model can ever be proved right, only wrong (*39,40*). If our observations are inconsistent with theoretical prediction, then we know something is amiss, but if our observations satisfy theoretical prediction, all we know is that the theory has not yet been proven wrong. Whether the theory will continue to work in the future is an open question. The longer a theory has been around and the more experimental tests it has passed, the more likely it seems that the theory is right but only in a probabilistic, not a deterministic, sense.

Scientists, of course, know this at least implicitly, and many modelers will argue that when they use the word validation they do not mean to imply that their model is literally true. They simply mean that it is not evidently false. The modelers have gone through a series of exercises to show that there are no major defects in the model and that they have done their "level best" (*41*). Validation, in this view, is a process of confidence building, of building a case for the model (*25,42,43*). A validated model, therefore, although not true strictly speaking, may be provisionally accepted (*44*). These are reasonable claims, hardly likely to provoke profound epistemic discontent, and they are certainly consistent with the first dictionary definition of the word valid: without obvious flaws or defects (*45*). From this definition, validation should simply imply the process in which obvious flaws are corrected.

But although these claims are reasonable, they are also problematic. One may remove obvious errors in a model while more subtle ones remain. If validation were merely the process of removing obvious defects, this would scarcely be sufficient for regulatory purposes. Regulatory agencies and the public seek assurance not merely that a model is free of gross error but that it provides a reliable basis for decision making (*19,20,46*). But to imply that the model provides a reliable basis for decision making is to imply that the model provides an accurate and substantially complete representation of the natural world. This, of course, is how people outside the modeling community interpret validation. In common usage, valid is taken as synonymous with correct, i.e., true, and elsewhere in the dictionary we find precisely that definition: "Valid implies being supported by objective truth" (*45*). The disclaimer that scientists know what they mean when they talk about validation would work if the models under discussion were being used solely within the confines of the relevant scientific communities. But very often they are not. Numerical simulation models are increasingly being used, often commissioned, by public agencies whose constituents are not privy to local scientific consensus.

Furthermore, individual scientists may claim that model validation does not imply an assertion about reality (*47*), but the official pronouncements of the regulatory agencies for whom they work frequently belie this claim. The Department of Energy, for example, has defined validation as the determination that a "model indeed reflects the behavior of the real world" (*48*). The International Atomic Energy Agency (*49*) has defined a validated model as one that provides a "good representation of the actual processes occurring in a real system." (The use of the word "actual" by the European agency is telling. In the 19th century, the French word actual was borrowed by both English and German scientists as a synonym for real and observable.) Protestations of scientists notwithstanding, it is evident why these regulatory agencies make these claims: Were they to describe validation only as a process of checking for gross error, it would be inadequate as a basis from which to forge political consensus (*50*).

A recent court case underscores this point. In 1986, the U.S. EPA was sued for failing to demonstrate the accuracy of a computer model used to set emissions limits under the Clean Air Act for two electric power plants in the state of Ohio. The question at stake was how much pollution could be emitted from the power plants without causing local air pollution levels to exceed federal standards, and the U.S. EPA had used a computer model to determine the answer. But the model was not predictively reliable. The resulting pollution levels violated the Clean Air Act, and the state government of Ohio took the U.S. EPA to court. The Sixth Circuit of the U.S. Court of Appeals ruled in favor of Ohio, finding against the U.S. EPA because it had used the computer model "without adequately validating, monitoring, or testing its reliability" (*51,52*). The U.S. EPA, the court concluded, acted arbitrarily in failing to establish the accuracy or trustworthiness of the model prior to basing decisions upon it, and ordered the agency "to test and validate the model as an adequate forecasting technique" (*53*). A notable feature of this case is that the utility companies that owned the plants were effectively shielded from liability for the pollution that their plants had caused because it was the U.S. EPA, not they, that had set the emission limits.

One could, of course, read this decision as implying that had the U.S. EPA validated the model, then the agency would have been blameless despite the model's predictive failure. After all, the action of the court was to order the U.S. EPA to validate the model! From this perspective, the more restricted notion of validation might at first sight appear adequate for regulatory purposes. But this is clearly not quite what the court intended. In the words of the decision, "In order to be useful, a model must accurately predict the 'behavior' of the...system being modeled." The argument of the petitioners against U.S. EPA was that "the model's predictions are not accurate..." (*53*). In fact, the U.S. EPA had validated the computer model: it had compared model output to empirical outcomes at four other sites. What the U.S. EPA had not done was test the model at the particular site and subsequently monitor the emissions. The court recognized that testing and monitoring at every site may not be practical—indeed, this is a primary reason for constructing simulation models in the first place—but it remains an open question as to how much site-specific testing and monitoring is required to satisfy legal and community standards. In this regard, scientists have an important role to play in openly discussing the problems and trade-offs involved.

Regulation and legal liability are not the only issues at stake here, nor are they, from a scientific and moral perspective, the most important ones. It may be possible to satisfy the legal standard of acting in a manner that is not arbitrary but fail to satisfy the scientific standard of producing reliable knowledge. Ultimately, the purpose of air pollution controls is to safeguard human health and property and preserve ecosystems. The purpose of the IEUBK model is to prevent new cases of lead poisoning. From this perspective, the issue is not whether the courts will be content with good-faith efforts, the issue is whether the model gives accurate results. In issues of public health and safety, we all have a stake in knowing that decisions made upon the basis of numerical simulation models turn out to be right.

Are Validated Models Valid?

Even if we were to set aside the conceptual issues raised by the example of celestial mechanics and accept the restricted definition of validation, i.e., that a valid model is one without obvious flaws or defects, would it then be possible to say that a given model is valid? The simple answer is no, because even our best models have known flaws. Science motivated by social needs may suffer this problem to a greater extent than science based on questions arising within a disciplinary framework. In the lab, scientists may define a problem in such a way as to rely primarily on areas where databases and conceptual understandings are very rich, and from this core of understanding venture outward toward the less well known. Scientists often refer to this as the well-posed problem. Throughout their history, scientists, both as individuals and as professional communities, have often set aside problems that could not be well posed.

Problems arising from social needs typically are not well posed because the world does not wait for scientific understanding. Where scientists have been asked to make models for use in policy domain, whether the issue is lead poisoning, global climate change, or the safe disposal of radioactive waste, our theoretical understanding and empirical databases are never what we wish them to be. There are always known flaws and defects in large, complex, policy-driven models.

We can think of these flaws as falling into four categories: theoretical, empirical, parametrical, and temporal. Theoretical flaws are the things we do not fully understand or do not have the mathematics to handle. In the case of lead toxicity, this would include, for example, the problem of differential susceptibility and the question of whether there is a safe threshold level of exposure. Empirical flaws are the things we cannot fully or precisely measure. This includes the pragmatic problem of having limited resources with which to measure lead in the environment, and the difficulties of sampling bias and analytical uncertainty, particularly at the very low exposure levels where regulatory limits will be set. Parametrical flaws are the errors introduced when we reduce complex empirical phenomena to single or simply varying input parameters in a model. Lead exposures vary continuously with time, for example, but models require input of a single value or a finite set of values for each individual. Likewise, blood lead levels are a continuously varying function, but we necessarily measure them at singular points in time, hoping that the points are adequately representative. Temporal errors arise from the assumption that systems are stable in time when they are not. For example, when we parameterize a lead model, we represent longitudinal lead exposure through cross-sectional lead measurements and assume, perhaps falsely, that these cross-sectional measurements are representative. Even if they are representative, it might be from a biologic standpoint that the highs and lows are as important as the means. Temporal variations may be important in ways that are neither fully understood nor even fully measured.

Validation versus Evaluation

Most scientists are aware of the limitations of their models, yet this private understanding contrasts the public use of affirmative language to describe model results. Published papers on validation are littered with positive terms: nouns like acceptance and substantiation, adjectives like satisfactory, adequate, and credible. The very word validation implies an affirmative result, that the process of validation will somehow validate the model (*32*). But where are the negative terms? If the purpose of validation is to determine whether a model is working well, shouldn't one also see nouns like rejection and refutation, adjectives like unsatisfactory and inadequate? The exercise of comparing a model with observations in the natural world is a test like any other scientific test, and it must be possible for a model to fail that test. If the context of validation is such that only positive results emerge, then something is wrong.

The conspicuous absence of negative language in the scientific literature of validation should give us pause, for it raises the following question relevant to both scientific and regulatory perspectives: Is the computer model a vehicle to prove what we think we already know or is it an honest attempt to find answers that are not predetermined? Put this way, it becomes clear that the goal of scientists working in a regulatory context should be not validation but evaluation, and where necessary, modification and even rejection. Evaluation implies an assessment in which both positive and negative results are possible, and where the grounds on which a model is declared good enough are clearly articulated. Validation implies an exercise in legitimation, and this is precisely what the public fears.

It is common to hear in regulatory and scientific circles that public fears are irrational, and there is substantial evidence that public fears are irrational if viewed from a statistical standpoint (*54,55*). But the language of validation does little to assuage such fears. Indeed, it exacerbates them because the public has learned (not without some justification) to be suspicious of reassurances (*6,55*). When citizens hear only positive claims, they start to doubt them, and they may sometimes be right: Some modelers have been guilty of exercises in legitimation of a predetermined result. A perhaps surprising example can be found in the work of the Club of Rome.

The world model was developed by Meadows et al. (*56*) in the early 1970s for the Club of Rome, a group of European industrialists, statesmen, and scientists concerned about overuse of natural resources. The model, described in the widely read book *The Limits to Growth*, predicted widespread natural resource shortages, exponential price increases for raw materials, and possibly global economic collapse before end of the century (*56*). The end of the century is here and resource use continues to grow, but proven reserves of natural materials are greater today than in 1972 and real prices are down for virtually all commodities (*57*).

One reason why the predictions of the world model have not come true is obvious in hindsight: the static way in which the model treated natural resources. Natural resources, such as copper, chromium, silver, and gold, were treated in the model as fixed and finite masses whose volumes could only decrease as use increased. On one level, this view is indisputable; the mass of chromium in the earth is a fixed (albeit unknown) number. But on another level, this view is hopelessly inadequate because it ignores the fact that the resource of chromium is not the same as the mass of it in the earth. A resource is something that may be used by humans. This involves a number of factors, including the price that people are willing to pay for it, the human and monetary capital available to look for it, the technology available for extracting it, and the cost of labor used to get it. A reserve is an even more constricted thing: reserves consist only of that portion of a resource that has been discovered, measured, and delineated.

The world modelers made an elision between the known reserves of a metal and the total mass of that metal in the world as if they were the same thing. But they are

not. Whereas the total mass of a metal in the earth must decrease or stay the same over time, reserves can increase as a result of increased exploration, improved technology, and/or decreased costs. Proven reserves of most metals have increased since 1973, primarily because of more and more effective geologic exploration during the past two decades, and prices have fallen as a result (*57,58*).

Why did the world modelers make what is in retrospect such an obvious mistake? One reason is revealed by the post hoc comments of Aurelio Peccei, one of the founders of the Club of Rome. The goal of the world model, Peccei explained in 1977, was to "put a message across," to build a vehicle to move the hearts and minds of men (*59,21*). The answer was predetermined by the belief systems of the modelers. They believed that natural resources were being taxed beyond the earth's capacity and their goal was to alert people to this state of affairs. The result was established before the model was ever built. In their sequel, *Beyond the Limits*, Meadows et al. (*60*) explicitly state that their goal is not to pose questions about economic systems, not to use their model in a question-driven framework, but to demonstrate the necessity of social change. "The ideas of limits, sustainability [and] sufficiency," they write, "are guides to a new world." (*60*)

One need not engage in an argument for or against social change to see the problem with this kind of approach if applied in a regulatory framework. The purpose of scientific work is not to demonstrate the need for social change (no matter how needed such change may be) but to answer questions about the natural world. The purpose of modeling is to pose and delineate the range of likely answers to "What if?" questions. The purpose of lead models should not be to demonstrate how bad lead ingestion is or how good U.S. EPA standards are but to try to find out what is most likely to happen if given standards are applied. The language of validation undermines this goal. It presupposes an affirmative result and implies that the model is on track. To outsiders, it raises the specter that the answer was preestablished.

There are other ways to talk about the problem. As Hodges and Dewar (*29*) write, the quality of a model is not equivalent to "agreement of the model with reality." Quality can be evaluated in several ways: on the basis of the underlying scientific principles, on the basis of quantity and quality of input parameters, and on the ability of a model to reproduce independent empirical data. All of these things can be discussed, but none of them should be discussed in either/or terms. Scientists should resist the demand to describe any model, no matter how good, as validated. Rather than talking about strategies for validation, we should be talking about means of evaluation.

That is not to say that language alone will solve our problems, or that the problems of model evaluation are primarily linguistic. The uncertainties inherent in large, complex models will not go away simply because we change the way we talk about them. But that is precisely the point: calling a model validated doesn't make it valid. The language of validation buries uncertainty; as scientists, we should be doing the opposite. We have an obligation to invite open discussion of uncertainties. And the more politically charged the issue at hand, the more essential it is that these uncertainties be articulated clearly, freely, and in language that anyone can understand.

One hundred years ago, Lord Kelvin famously tried to eliminate uncertainty over the age of the earth. Based on the concept of uniformitarianism, the assumption that observable geologic processes are representative of earth history in general, geologists in the late 19th century concluded that the earth was probably a few billion years old. But they had no way to prove it, and efforts to calculate the earth's age precisely had produced numbers as low as 100 million and as high as several hundreds of billions. Kelvin, famous for his penchant for quantitative precision, applied Fourier's theorem of conductive cooling to the question. Assuming that the earth has solidified from an incandescent globe, he obtained a maximum time of 98 million years for it to have cooled to its present surface temperature, and he promptly declared the entire science of geology invalid. Any conceptual scheme that implied a billion-year old earth was fundamentally flawed, he declared. Pursuing the same logic, he dismissed Darwin's theory of natural selection on the grounds of inadequate time for it to operate (*61,62*).

For several decades, Kelvin's more certain result held sway and evolutionists were in nearly full retreat until the discovery of radiogenic heat proved that it was Kelvin rather than the geologists whose conceptualization was faulty. We know now, of course, that the earth is 4.5 billion years old, more than enough time for natural selection to have operated as Darwin envisaged it. Kelvin's calculations, although theoretically valid and highly precise, produced a result inaccurate by a factor of 50. In his desire for certainty, Lord Kelvin made one of the most colossal blunders in the history of modern science. As his infamous mistake clearly shows, the uncontrolled desire for certainty may lead to fallacious quantification and a false sense of security.

REFERENCES AND NOTES

1. Rosner D, Markowitz G. A "gift of God"? The public health controversy over leaded gasoline in the 1920s. Am J Public Health 75:344–352 (1985).
2. Lippmann M. 1989 Alice Hamilton Lecture. Lead and human health: background and recent findings. Environ Res 51:1–24 (1990).
3. Mushak P. Defining lead as the premiere environmental health issue for children in America: criteria and their quantitative application. Environ Res 59:281–309 (1992).
4. Brush SG. Transmuted Past: The Age of the Earth and the Evolution of the Elements from Lyell to Patterson. Cambridge, UK:Cambridge University Press, 1996.
5. Ducoffre G, Claeys F, Bruaux P. Lowering time trend of blood lead levels in Belgium since 1978. Environ Res 51:25–34 (1990).
6. Nelkin D, ed. Controversy: Politics of Technical Decisions, 3rd ed. Newbury Park:Sage Publications, 1992.
7. Bornschein RL, Clark CS, Grote J, Peace B, Roda S, Succop P. Soil lead-blood lead relationship in a former lead mining town. In: Lead in Soil: Issues and Guidelines (Davies BE, Wixson BG, eds). Northwood, UK:Science Reviews, 1988;149–160.
8. Miller GD, Massaro TF, Massaro EJ. Interactions between lead and essential elements, a review. Neurotoxicology 11:99–120 (1990).
9. O'Flaherty EJ. Physiologically based models for bone seeking elements. IV: Kinetics of lead disposition in humans. Toxicol Appl Pharmacol 118:16–29 (1993).

10. O'Flaherty EJ. Physiologically based models for bone-seeking elements. V: Lead absorption and disposition in childhood. Toxicol Appl Pharmacol 131:297–308 (1995).
11. Menton RG, Burgoon DA, Marcus AH. Pathways of lead contamination for the Brigham and Women's Hospital Longitudinal Lead Study. In: Lead in Paint, Soil and Dust: Health Risks, Exposure Studies, Control Measures, Measurement Methods, and Quality Assurance. ASTM STP 1226 (Beard ME, Allen Iske SD, eds). Philadelphia:American Society for Testing and Materials, 1995;92–106.
12. U.S. EPA. Air Quality Criteria for Lead. EPA-600/8-83-028. Research Triangle Park, NC:U.S. Environmental Protection Agency, 1986.
13. Davis JM, Svendsgaard DJ. Lead and child development. Nature 329:297–300 (1987).
14. Leggett RW. An age-specific kinetic model of lead metabolism in humans. Environ Health Perspect 101:598–616 (1993).
15. U.S. EPA. Validation Strategy for the Integrated Exposure Uptake Biokinetic Model for Lead in Children. EPA 540/R-94-039. Washington:U.S. Environmental Protection Agency, 1994.
16. Marcus AH, Cohen J. Modeling the blood lead - soil lead relationship. In: Lead in Soil: Issues and Guidelines (Davies BE, Wixson BG, eds). Northwood, UK:Science Reviews, 1988; 161–174.
17. Marcus AH, Elias RW. Estimating the contribution of lead-based paint to soil lead, dust lead, and childhood blood lead. In: Lead in Paint, Soil and Dust: Health Risks, Exposure Studies, Control Measures, Measurement Methods, and Quality Assurance. ASTM STP 1226 (Beard ME, Allen Iske SD, eds). Philadelphia:American Society for Testing and Materials, 1995;12–23.
18. Hogan KA, Elias RW, Marcus AH, White PD. Unpublished data.
19. Jasanoff S. The misrule of law at OSHA. In: The Language of Risk: Conflicting Perspectives on Occupational Health (Nelkin D, ed). Beverly Hills, CA:Sage Publications, 1985;155–178.
20. Jasanoff S. Science, politics, and the renegotiation of expertise at EPA. Osiris 7:195–217 (1992).
21. Shakley S. Trust in models? The mediating and transformative role of computer models in environmental discourse. In: International Handbook of Environmental Sociology (Redclift M, Woodgate G, eds). (Forthcoming). Cheltnham, UK: Edward Elgar, 1997; 237–260.
22. Marcus AH, Elias RW. Model Validation Workshop. Draft Outline. Research Triangle Park, NC:U.S. Environmental Protection Agency, 1996.
23. Balls M, Blaauboer B, Brusick D, Frazier J, Lamb D, Pemberton M, Reinhardt CA, Roberfroid M, Rosenkranz H, Schmid B, et al. Report and Recommendations of the CAAT/ERGATT Workshop on the Validation of Toxicity of Test Procedures. ATLA Altern Lab Anim 18:313–227 (1990).
24. Balls M, Blaauboer B, Fentem JH, Bruner L, Combes RD, Ekwall B, Fielder RJ, Guillouzo A, Lewis RW, Lovell DP, et al. Practical aspects of the validation of toxicity test procedures: report and recommendations of ECVAM Workshop 5. ATLA Altern Lab Anim 23:129–147 (1995).
25. Dee DP. A pragmatic approach to model validation. In: Quantitative Skill Assessment for Coastal Ocean Models (Lynch DR, Davies AM, eds). Washington:American Geophysical Union, 1994.
26. Curren RD, Southee JA, Speilmann H, Liebsch M, Fentem JH, Balls M. The role of prevalidation in the development, validation and acceptance of alternative methods. ATLA Altern Lab Anim 23:211–217 (1995).
27. Bruner LH, Carr GJ, Chamberlain M, Curren RD. Validation of alternative methods for toxicity testing. Toxicol in Vitro 10:479–501 (1996).
28. Marcus AH, Elias RW. Some useful statistical methods for model validation. Environ Health Perspect 106(Suppl 6). 1541–1550 (1998).
29. Hodges JS, Dewar JA. Is It You or Your Model Talking? A Framework For Model Validation. Rpt no. R-4114-A/AF/OSD. Santa Monica, CA:RAND Corporation, 1992.
30. Bankes SC. Exploratory Modeling and the Use of Simulation for Policy Analysis. Rpt no. N-3093-A. Santa Monica, CA:RAND Corporation, 1992.
31. Hodges JS. Six (or so) things you can do with a bad model. Operations Res 39:355–365 (1991).
32. Oreskes N, Shrader-Frechette K, Belitz K. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263:641–646 (1994).
33. Oreskes N. Testing models of natural systems: can it be done? In: Structures and Norms in Science, (Chiara ML, Doets K, Mundici D, van Benthem J, eds). Amsterdam:Kluwer Academic Publishers, 1997;207–217.
34. Van Damme K, Casteleyn L, Heseltine E, Huici A, Sorsa, M, van Larebeke N, Vineis P. Individual susceptibility and prevention of occupational diseases: scientific and ethical issues. J Occup Environ Med 37:91–99 (1995).
35. Konikow LF, Bredeheoft, JD. Ground-water models cannot be validated. Adv Water Resour 15:75-83 (1992).
36. Nordstrom DK. On evaluating and applying aqueous geochemical models. EOS Trans Am Geophys Union Suppl (April 20):326 (1993).
37. Kuhn TS. The Copernican Revolution: Planetary Astronomy in the Development of Western Thought. Cambridge, MA:Harvard University Press, 1957.
38. Hempel CG. Philosophy of Natural Science. Englewood Cliffs, NJ:Prentice-Hall, 1966;23–24.
39. Popper KR. The Logic of Scientific Discovery. New York:Harper Torchbooks, 1937.
40. Popper, KR. Conjectures and Refutations: The Growth of Scientific Knowledge. New York:Harper Torchbooks, 1963.
41. de Marsily G, Combes P, Goblet P. Comment on "Groundwater models cannot be validated" by L.F. Konikow and J.D. Bredehoeft. Adv Water Resour 15:367–369 (1992).
42. Neuman SP. Validation of safety assessment models as a process of scientific and public confidence building. In: High Level Radioactive Waste Management: Proceedings of the Third International Conference, 12-16 April 1992, Las Vegas, Nevada. New York:American Society of Nuclear Engineers, 1992;1404–1413.
43. Nir A, Doughty C, Tsang DF. Validation of design procedure and performance modeling of a heat and fluid transport field experiment in the unsaturated zone. Adv Water Res 15:153–166 (1992).
44. Rykiel EJ. The meaning of models [Letter]. Science 264:330–331 (1994).
45. Woolf HB, ed. Webster's New Collegiate Dictionary. Springfield, MA:G&C Merriam Co, 1973.
46. Beer FA. Validities: a political science perspective. Soc Epistemol 7:85–105 (1993).
47. Younker JL, Boak JM. Geological models [Letter]. Science 264:1065 (1994).
48. U.S. Department of Energy. Environmental Assessment Overview: Yucca Mountain Site, Nevada Research and Development Area. Rpt no. DOE/RW-0079. Washington: Office of Civilian Radioactive Waste Management, 1986.
49. International Atomic Energy Agency. Radioactive Waste Management Glossary. Doc no IAEA-TECDOC-264. Vienna:International Atomic Energy Agency, 1982.
50. Jasanoff S. The Fifth Branch: Science Advisors as Policymakers. Cambridge, MA:Harvard University Press, 1990.
51. Ohio vs U.S. Environmental Protection Agency. U.S. Law Week 54:2494 (1986).
52. Davis PA, Olague NE, Goodrich MT. Approaches for the Validation of Models Used for Performance Assessment of High-Level Nuclear Waste Repositories. SAND90-0575/NUREG CR-5537. Albuquerque, NM:Sandia National Laboratories, 1991.
53. Ohio vs U.S. Environmental Protection Agency. U.S. Court of Appeals, Sixth Circuit, 23 ERC 2091-2097, 1986.
54. Cohen BL. Before It's Too Late: A Scientist's Case for Nuclear Energy. New York:Plenum Press, 1983.

55. Shrader-Frechette KS. Risk and Rationality. Los Angeles:University of California Press, 1991.
56. Meadows DH, Meadows DL, Randers J. The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. New York:Universe Books, 1972.
57. Simon JL, Kahn H, eds. The Resourceful Earth: A Response to Global 2000. Oxford:Blackwell 1984.
58. Tierney J. Betting the planet. New York Times Magazine (December 2):52–81 (1992).
59. Peccei A. The Human Quality. Oxford:Pergamon Press, 1977.
60. Meadows DH, Meadows DL, Randers J. Beyond the Limits: Confronting Global Collapse, Envisioning a Sustainable Future. White River Junction, VT:Chelsea Green Publishing Company, 1992.
61. Burchfield, JD. Lord Kelvin and the Age of the Earth. Chicago:University of Chicago Press, 1990.
62. Smith C, Wise MN. Energy and Empire: A Biographical Study of Lord Kelvin. Cambridge, UK:Cambridge University Press, 1989.