JOURNAL OF COMPUTATIONAL PHYSICS 21, 245–250 (1976)

Letter to the Editor

Failure Testing: A Proposal for Increasing Confidence in the Results of Numerical Simulations

> "What then can we look for from this confusion of groundless and extraordinary opinions but error and falsehood? And how can we justify to ourselves any belief we repose in them?"
>
> David Hume, *A Treatise of Human Nature*

The maturing of the techniques of numerical simulation over the past 20 years has been marked by an essential change of outlook. Computer calculations, which were initially viewed as extensions of analytical studies and approximations into realms too complex or too extensive to be performed by hand, have now in many applications attained the status of actual experiments. If in the laboratory certain physical parameters can be fixed and the behavior of others studied, the same can be accomplished in the complex numerical model of a simulation code. Often the numerical experiment can be performed quicker or cheaper than the real one; occasionally the real experiment would itself be impossible in a terrestrial laboratory.

But if these simulations are replacements for or approximations to experiments that have not been performed, how can we know that the results of simulation and experiment would agree, that, in short, the simulation means anything in terms of the real physical world? Some fixed, uniform standards should exist which could provide us with at least a guide to or rough measure of the reliability of numerical simulations. But such standards do not exist; only the standards for mathematical approximations have even been seriously investigated. The purpose of this letter is to elucidate the source of the great difficulty in the formation of standards and to suggest a route by which they might begin to be formulated.

1. The Rationality of Computer Simulation

Numerical simulation and physical experiment differ in two essential ways. In one sense the simulation is more limited, because it is capable of dealing only with principles supplied *ab initio*, and hence presumed known; the experiment may contain entirely unknown phenomena. But in a second sense we presume (else

there is no purpose to experiment at all) that the experiment in "querying nature" is sampling a fixed trustworthy logical structure, which is merely imperfectly known; in performing a numerical simulation the scientist is querying his hypotheses themselves, testing them for a logical consistency which is by no means guaranteed.

Of course our inability to discover through simulation phenomena beyond known principles does not mean that the process is useless, since many unknown phenomena still exist within known principles. But we must be aware of the unbounded nature of the logical order with which we are dealing. It should always strike us as peculiar that one may reach a solution to a physical problem by purely mathematical means. In arriving at what he can verify as a correct solution to a control problem, the physicist may appeal to mathematical logic whose effect and nuances he is perfectly acquainted with, yet still for a time not understand fully the real physics underlying the appropriateness of his mathematical choices. This is partially a result of the fact that one is constrained to operate within a small, restrictive set of assumptions about the physics; it is far easier for the results of mathematical manipulations to be clear in the artificial system, where only a few equations pertain, than for the effects of excluded physics (and their interactions with included physics) to be evident in the same system, where everything that is known is by definition approximate. Thus the real world is more complex first than *any* conception of it, and second than any logical structure. This is why nature can be self-consistent in spite of the fact that we cannot demonstrate that any system of mathematics is, for mathematics is but one further construct.

The crux of the problem is that, absolute verification by comparison with experiment being impossible, no amount of approximate verification (by means of "test problems" with known solutions) will suffice to demonstrate a simulation's verisimilitude with the physical world. The difficulty is in essence caused by the failure of inductive logic. David Hume's famous criticism demonstrated that inductive logic cannot be rigorous in the sense of compelling a conclusion; no matter how many times we may observe a particular congruence, we can never be justified in concluding solely on the grounds of that previous knowledge that apparently identical circumstances will produce the same result. This conclusion does not depend upon a presumtion of the possibility of unknown forces at work, but applies to the logical process of surmising based upon past events. We may experience a greatly enhanced confidence in our generalizations, but a certainty is impossible.

Therefore no combination of successful test problems can guarantee that when a particular code is applied to imperfectly understood phenomena (that is, employed to a useful rather than a trivial purpose) the results will resemble the true phenomena. Accuracy in simulations of test problems may serve to increase confidence in the code, but it can be cogently maintained that in adjusting the code

to perform tests properly one is in fact tailoring to a more specific rather than more general reality. Further, a danger of this sort of test by *a posteriori* verification is that one may be creating a model with no unambiguous content, in the same same sense that psychiatric theory is capable of self-modification in a continuous flow to account for any empirical data.

Test problems, then, produce an illusory confidence. And no matter how carefully done they tell us little concerning the question of limitations of the model: Precisely what range of physics is or is not represented? In a real experiment we may be at pains to learn whether a particular model is or is not equivalent to a particular phenomenon, and in fact we may not be able to say for certain without further knowledge, but we can at the very least make an assignment of class (does represent, does not represent, cannot know) on the basis of information at hand. Such an assignment gives us valuable advice on which direction to proceed in search of further information. But in numerical simulation we often cannot make such an assignment with confidence: a sort of Gödel's law is in play. In some instances we cannot be certain that unknown artificial (numerical, logical, or modeling) effects are dominant, and so we cannot say whether or not the code is adequate, or even whether we cannot say if it is. Hume's dictum indicates that in some classes this uncertainty may be absolute, and the only thing we can say is that we cannot say.[1]

Hence there are two general classes of uncertainties related to verification by test problems, one arising from the practice itself and one which it simply fails to address, that make test problems insufficient as a means of arriving at high levels of confidence in numerical simulations. Tests are enormously helpful in eliminating most fundamental errors from a code, and certainly increase confidence in the simulation of analytically approachable phenomena. But there still remain regions of pressing import into which their influence does not extend.

2. The Elusive Trustworthy Code

The degree to which Hume's criticism and its implications reduce all science from rational certainty to relative confidence is still a matter of extensive debate

[1] If the analogy with Gödel's law holds fully, of course, we shall have the added disadvantage of never being able to demonstrate that for many particular difficult problems our solution is adequate even when we are quite certain that it is. Thus a "second degree" of uncertainty is established, in which even an assignment into a class is formally meaningless. Fortunately, we seldom care to follow the analysis this far. Attempting to formulate such a dictum rigorously would lead us to the creation of a frustrating third-degree of uncertainty and the waste of much time far from the physical problems; it would also bring us up against the paradox that such a determination should *also* be forbidden by Gödel's law (an absolute test for nonapplicability would lead to an absolute criterion of applicability).

among scientists and philosophers. Quite beyond that debate is the faith that all working scientists must have, that logical approaches will open a pathway to physical truth. Within the logical framework, it is clear that most physicists and virtually all applied physicists do rely for a base of confidence on a current paradigm of physics: Though they may be aware that difficulties in the philosophical footing exist, most physicists daily apply with confidence the laws of mechanics and optics and thermodynamics.

Yet, curiously, for the computational physicist working wholly within the paradigm, the logical uncertainties can never be ignored and should remain present in his working and thinking environment. For confidence, as we discussed, is always a pressing question in numerical simulations; and in the explanation of the behavior of every simulation the fallacy of the excluded middle term looms suggestively in the background.

Conceding that science cannot attain absolute truth, the philosopher Karl Popper [1,2] has suggested a justification for the rational authority of science, a viewpoint from which science does remain autonomous, based entirely upon observation and logic. Popper views the progress of science as a series of conjectures (hypotheses) concerning the unknown which make definite predictions, predictions which may be refuted by further observation or experiment. The greater the risk of refutation (the more easily or directly it may be refuted), the more meaningful the hypothesis. To Popper, supportive and interpretive evidence are meaningless as science, for theories constructed from them, such as astrology or Marxism, may never risk refutation, simply because categorical refutation may be impossible. Though a particular theory may never attain absolute truth, we may come to place very high confidence in it through its repeated exposure to the risk of refutation.

This risk-of-refutation concept can provide the basis for methods of establishing confidence in the validity of numerical simulations. The simplest process fully embodying this concept is one which I call failure testing. Typically, numerical simulations are characterized by three different types of parameters: (1) those which represent physically real quantities being studied, such as the mass, density or temperature of a fluid; (2) those which are purely artifacts of the numerical approximations to the equations defining the system and its interactions, such as timestep or grid interval or macroparticle shape; and (3) those which are physical quantities represented in a nonphysical approximation, such as an average charge per particle, an imperfectly transmittive boundary or a too perfectly conducting boundary. The one thing that all of these parameters have in common is that they are easily available for adjustment. Of course, ease of adjustment and evaluation of physical parameters is one of the primary familiar advantages of numerical simulation over real physical experiments. Less often mentioned is the comforting fact that when the value of a parameter is incorrectly or injudiciously assigned, the experiment is not destroyed; all that is lost is a little computer time. This

feature suggests further interpretation that would permit rigorous tests of refutability of predictions.

If one were *intentionally* to set the values of his code's parameters, in all three categories, at the border of and well into forbidden regimes, one should have been able to predict beforehand with considerable accuracy how the code would fail as a representation of physical phenomena or what new artificial phenomena would be introduced.[2] This is the notion of failure testing; it involves, in effect, running experiments of a secondary nature to question the reality and self-consistency of a code and the models it is constructed from. The critical element in the procedure is the formulation of a precise prediction of the failure threshold and mode, the exact nonphysical results, in a form which can be categorically refuted if incorrect. The results of this process do not guarantee a logically rigorous solution any more than a physical experiment may in Popper's philosophical construct, but they approach statistical confidence in predictive capability and verisimilitude in precisely the same fashion.

A thoroughly persued program of failure testing (in which each identificable parameter is tested at its reasonable limits) would fill many of the logical gaps left by the process of running test problems. Both reliability and self-consistency are verified, and the nagging question of tailoring one's code to one's samples is done away with. Though problems of excluded physics are not entirely met, two important subsets are taken care of: limits can be set on classes of included physics which are significant to the code, thus establishing clearly those which are not; and one may eliminate a large fraction of the numerical interactions which masquerade as physical phenomena (such as the celebrated numerical Cherenkov instability in electromagnetic particle codes [3]).[3]

Now a thoroughly done job of failure testing of a new code (or new set of

[2] Trivial examples might help to clarify the procedure. For instance, one might, in a plasma particle simulation code, raise the timestep past the plasma period to induce instabilities; in a fluid code one might raise viscosity to remove detailed motion artificially; in an MHD code one might permit the time step to violate the Alfvén-speed Courant condition. Of course, one would expect virtually any code to respond well to such simple tests, but they could verify methodology quickly. More complex tests could easily be chosen for particular new codes.

[3] A more relevant and striking example is provided by the history of the electron–cyclotron drift instability. The nonlinear behavior of this instability, inaccessible to analytical mathematical techniques, and probably beyond sufficiently controlled laboratory procedures, was finally settled through computational studies [4]. But the demonstrated saturation mode depended upon destruction of the integrity of cyclotron orbits caused by extreme fluctuations in electrostatic fields; and the severest problem in simulations of the kind used to investigate this instability is high-frequency electric noise fundamental to the computational approximations. Settling of the dispute over the role played by such "artificial" effects was the occasion of a long and bitter dispute among highly competent colleagues. Very fine mathematical points are capable of introducing what appear to be real physical phenomena into a simulation, and identifying these demons at an early stage can be crucial to the usefulness of a code.

simulations with an old code) would, admittedly, be a lot of trouble to go to; any thoroughly done job usually is. Yet there is scarcely a computational physicist who has not had the (often unpleasant) experience of mispunching an input card only to find that, having discovered his mistake, he still cannot account for the results on his desk. For that matter, there are fewer still who have not had the experience of printing or plotting a hitherto unexamined variable in a well-trusted code, only to discover symmetries or asymmetries they cannot explain. Thoroughness in examining and interpreting all one's range of available data has been argued often; though highly regarded as a professional practice, it is seldom done. Perhaps there is little hope for reliable failure testing, then (of which such full data interpretation may be seen as a preliminary, even an included process). There is some hope in the fact that the full process can be streamlined by carefully considering other things one knows about code, including the results of other tests, of analysis and of experience: certain mathematically rigorous conclusions may be appealed to, such as for noise levels in working finite Fourier transforms; and experience and theory may set some combined limits, such as for the potential noisiness of certain combinations of grid sizes and particle densities.

Still, in the real world the amounts of pre- and post-analysis one performs are limited principally by working time and expense. The questions of credibility and confidence must be accepted as critical ones, and attacked in some rigorous methodology, before numerical simulation ceases to be regarded by analytical and experimental physicists alike as a black art.

REFERENCES

1. K. R. POPPER, "The Logic of Scientific Discovery," Basic Books, New York, 1961.
2. K. R. POPPER, "Conjectures and Refutations: The Growth of Scientific Knowledge," Harper Torch Books, New York, 1968.
3. B. B. GODFREY, *J. Comput. Phys.* **15** (1974), 504.
4. M. LAMPE, *et al.*, "Nonlinear Development of the Beam Cyclotron Instability," NRL Memorandum Report 2238, Naval Research Laboratory, Wash., D.C., April 1971.

RECEIVED: October 23, 1975

THOMAS H. JOHNSON[4]

Department of Applied Science,
Davis/Livermore,
California

[4] Present address: Air Force Weapons Laboratory, Kirtland Air Force Base, New Mexico 87117.

Printed in Belgium