Abstract. In an era of accountability in education, how do we assess student understanding of nature of science (NOS)? I profile an alternative to VNOS and similar assessment instruments that are based on short lists of declarative NOS elements. Instead, our aim is to assess skills in analytical and critical thinking, fundamental to scientific literacy. One prototype probes whether a student can produce a well informed analysis of contemporary case studies. This form of assessment invites NOS instruction based on case studies designed as samples of 'whole science'.

 Doubting everything or believing everything are two equally accommodating solutions, either of which saves us from reflection.  

–Henri Poincaré, *La Science et l'Hypothèse* 

Given the special character of this group, I have endeavored, in earlier presentations — on myth-conceptions and lawless science — to find provocative perspectives that might stimulate fruitful reflection beyond familiar conceptual boundaries. I wonder sometimes if, by questioning certain assumptions, my views have been dismissed as too critical or unrealistic.

Today, I will continue the pioneering spirit, but with a conspicuously concrete and practical posture, as exhibited in my work writing case studies (in *Doing Biology*) and developing resources (on the SHiPS website). My focus is a critical problem that has remained largely unsolved for several decades: evaluating nature-of-science knowledge.

The challenge is framed in part by the current politics of education, at least in the U.S., with its emphasis on accountability. Realistically, teachers have few options. They inevitably teach to the test. Practicing teachers feel that they don't have enough time to teach what's important, as documented in this year's extensive study by the Gates Foundation. Clough and Olson noted, too, in introducing a special 2008 journal issue on nature of science, "efforts are urgently needed to help teachers and creators of high stakes tests to accurately assess
students’ understanding of NOS." As David has just sketched, VNOS is flawed and unacceptable. We need to develop an effective NOS test — and write it in such a way that we may endorse teachers teaching to it. Write it, adopt it, and they will teach to it.

My benchmark is the aim of scientific literacy [news]. Inflated rhetoric aside, a foundational touchstone may be simply stated:

Students should develop an understanding of how science works with the goal of interpreting the reliability of scientific claims in personal and public decision-making.

This practical, public goal was echoed again earlier this year by Krajik and Sutherland in Science magazine. An approach that emphasizes the context of personal and public decision-making gives central significance to analysis of reliability, trustworthiness or authority. What students need to learn foremost is what, or who, to trust — and why. Such assessments are especially important where scientific knowledge is "young" — that is, has not yet withstood the proverbial "test of time": the status of many if not most scientific claims in contemporary cases. Public science education is foremost about preparing students to generate a well informed analysis in such cases from the recent news as Climategate, mammogram recommendations, facilitated communication of coma patients, and autism and the measles vaccine. To interpret such cases, one may need to understand some basic scientific concepts as background or be able to assess simple evidence. But for many cases, understanding the nature of science instead is essential, if not central: Whose expertise can be trusted, especially when experts seem to disagree? What forms of communicating scientific findings to the public are credible? How do scientists manage data? How do they communicate with each other? What kind of conditions warrant a change in scientific consensus? Where does verifiable information end and value judgment or ideology begin? How might scientists make "honest" mistakes, and how does one detect them? These are the important questions that a
prospective test must address. Not: "what is an experiment?", "Are scientists creative?", or "How do laws and theories differ?"

Let me skip the tedious analysis (but feel free to ask!) and present a concrete prototype, modeled on the AP exam essay and its scoring format. The question is based on a case from current events, as a typical citizen might encounter it in the news media — here the revised mammogram recommendations issued in the U.S. last November:

A female acquaintance of yours is just turning 40. Concerned about the possibility of breast cancer, she had planned to get a mammogram in the next few months, despite her fears about excessive radiation. She has heard that a major national Task Force now advises waiting until 50, yet finds reassurances in Women's Health magazine about still following the old guidelines. You both knew another woman who was diagnosed unexpectedly with breast cancer at age 43 and died last year. Your acquaintance is unsure how to interpret the apparently conflicting advice and asks your view. Provide an analysis of this reported change in scientific consensus that would help inform her decision.

Students are provided several documents as a resource: from Women's Health magazine, The New York Times, the original U.S. Preventative Services Task Force report, and an editorial from the Annals of Internal Medicine — all available on the internet.

① First, this approach focuses on analyzing extended examples or cases. Such was the informal consensus of a group of about two dozen philosophers of science, science policy scholars, high school science teachers and scientists, who discussed this problem last summer at the Second Biennial Meeting of the Society for the Philosophy of Science in Practice. Participants commented on the need to allow students to articulate NOS understanding in a case that is sufficiently complex, on the one hand, yet also clearly concrete and delineated on the other. This view paralleled another consensus: that NOS instruction
likewise needs to engage students in problem-solving in context-rich case studies. Everyone agreed that students need experience and guidance, possibly in simplified contexts, for such real-life scenarios. They also implicitly endorsed the notion that if scientific literacy is the educational goal, scientific-literacy-in-practice should guide both assessment and instruction. That is, both should be authentic, widely acknowledged now as an evaluation ideal. More on the teaching dimension later.

I have developed similar prototype questions on Climategate, autism and the measles vaccine, facilitated communication of coma patients and beriberi in Java in the 1890s, often asking the student to adopt a specific perspective — researcher, policy-maker, citizen-advisor, or consumer. All are now available online, along with the supplemental public documents: news items, blogs, position statements, government publications, etc. More on historical cases in a moment.

2 Second, these questions seek a well informed analysis. They do not ask students "What is your view?" or "What would you do?" Asking for a particular position, or worse, a justification, only tends to elicit post hoc rationalizations. Any assessment must avoid values, ideology and personal judgment. Rather, it should probe how well a student understands scientific practice in ways relevant to decision-making. As recommended in a 2009 U.S. National Academy of Sciences report, the aim is to reveal the student's thinking. In particular, it asks them to analyze NOS features, independent of specific content knowledge. Students are not clinical researchers able to assess mammogram effectiveness on their own. The question focuses instead on how the trustworthiness of the new recommendations relate to NOS features: credibility of sources, the nature of conceptual change, emotional bias in interpreting evidence and risk, the role of prior beliefs, the role of systematic study versus anecdote, sample size, the role of probability in inference, gender bias, sources of funding. One might also ask the student identify information that would be critically important, but is
missing—and where one would likely secure such information. Ultimately, the test is about analytical skills and competence, not declarative knowledge. This, too, echoes emerging norms on assessment.

③ Third, the question format can function in a formal mass-testing setting, but also in a project- or performance-based portfolio assessment, yet another trend. There need not be some universal "test day." Teachers can assess their own students in local settings, even if the questions are standardized and calibrated on a wider scale. Students may complete a series of such questions, guided by the local teacher until they can demonstrate competence working independently: the ultimate goal.

④ Fourth, the question assesses both breadth and depth of NOS knowledge. For example, does the student identify and articulate the significance of the relevant NOS factors from the information provided? Can the students discern practices that foster or inhibit reliable conclusions? The format allows for both qualitative feedback as well as quantitative scoring — allowing for both formative and summative assessment. Through simple scoring, it also allows for comparison across individuals, teachers, schools, districts, states, satisfying those focused on such metrics. The flexibility to accommodate multiple assessment styles — both formative and summative, both test-based and project-based, both student-centered and comparative— is a significant strength of this style of question, and thus may be applied widely.

Now, for the practicalities of scoring. An appropriate answer will be free response, or essay. This format has been extensively developed by the Educational Testing Service for their Advanced Placement (AP) exam, whose scoring is routinely standardized. Indeed, as I noted, I have used the AP essay as a guide to practicality. Having participated in grading exam that for over ten years, I am ready to answer questions about the process. A sample scoring rubric here will work from a limited set of possible points — no more than 10 (one for each finger!).
Consider that a student earns points for effectively identifying and articulating the significance of each relevant NOS factor, as enumerated earlier: 1 point. With elaboration, such as informative comparison to other cases, another point. Again, the aim is well informed analysis, not particular answers. Mere statements of general principles, without concrete examples or context — for example, saying only that "all scientific knowledge is provisional" — may not be sufficient. In a summative context, one needs only minimal ranking — say, across 4 or 5 levels from "inadequate progress" to "exemplary" — so three or four questions on a ten-point scale will likely suffice to monitor student performance or progress across a cross-sectional sample of NOS features.

Now, the complementary task for educators and science studies scholars — no less challenging — is to articulate all the NOS dimensions in sufficient detail, at least within the context of assessing reliability of particular knowledge claims. Here is one catalog, as a starting point. It identifies important methods that help ensure reliable conclusions — namely, all the occasions where error can emerge if the scientist falters. It basically lists sources of error, or error types, in science — and thus the corresponding principles for reliability. The list spans experimental, cognitive and socio-cultural dimensions. It ranges from relatively narrow, or local claims relevant in a laboratory setting, to broader, or more global claims synthesizing large amounts of observations and data. From test tubes to YouTube. How does each factor contribute to claims one can trust? All are addressed in the Science Studies literature. Effective NOS instruction should thus help students develop analytical skills in each category. Effectiveness will be measured not by agreement with prescribed statements, but by the degree, or depth, to which a student is informed about the factors that shape reliability in each dimension. Here, NOS knowledge has both breadth and depth — completeness and proficiency — and each are assessed through the prototype questions.
Now, on to teaching — and more on historical cases. Assuming these questions capture a sense of the goal of teaching scientific literacy — and others tell me that they do—one may work "backwards" to the context of teaching. According to a recent U.S. National Academy of Sciences study, assessment formats are ideally modeled in previous classroom activities and exercises, thereby allowing teachers the opportunity to guide students towards appropriate target performance. That means — no surprise to anyone here — introducing students to complex, or fully contextualized, case studies. Plural, not singular. This was certainly the view in the landmark 2003 study by Osborne, Collins, Ratcliffe, Millar and Duschl on what 'ideas-about-science' should be taught in school science:

Ideas about science, are perhaps best . . . addressed [they claimed] through sets of well-chosen case studies of either a historical or contemporary nature and by more explicit reflection and discussion of science and its nature—an aspect that should emerge naturally from the process of scientific inquiry that is a normal feature of much classroom practice. (p. 716)

The prospective assessment, however, underscores the additional importance of engaging students in solving problems. Problem-based learning, of course, is a major trend in education at many levels. Narratives, certainly effective as a presentational format and contextualizing strategy, should ideally introduce problems. And some of those problems should be NOS problems: that is, NOS problematized for students to analyze or resolve in an authentic context. A case study becomes not just an occasion to tell a story, or deliver a pre-digested lesson, but rather a scenario to engage students in active learning about nature of science. In particular, cases should not just memorialize heroes or promote particular philosophical views. Literacy is about understanding how science works, not embracing some private ideology. We must safeguard against values masquerading as the nature of science.

Now where, you may be asking, where is the history in all this? [A show of hands,
please?] Doesn't this emphasis on contemporary cases tend to eclipse a role for history?

Absolutely not. Modern cases lack one feature critical in an instructional context: a clear solution by which to judge one's emerging problem-solving or interpretive efforts. Students need the freedom to fail while they practice new analytical skills. They also need to understand how the outcome ultimately unfolded as a benchmark to evaluate and adjust their own maturing skills. Consider how a student's interpretation of the mammogram case would be informed by learning about how more data unexpectedly shifted consensus on the causes of pellagra or beriberi or the risk assessments of thalidomide and genetic engineering, or about popular views of phrenology.

In addition, history seems essential for conveying certain lessons about the nature of science, which rely on retrospect or larger perspectives. These include, most notably: the ways scientists can err, the nature of conceptual change and uncertainty, and the role of cultural context and potential bias in scientific ideas. ‘Tentativeness’, or the provisional nature, of scientific knowledge, in particular, has been a prominent NOS learning goal for decades. It is also a basic epistemological belief fundamental to the ability to learn on one's own beyond the classroom. To teach this, one needs examples of real, profound, and unanticipated conceptual change. To enable informative contrast of a reasonable "before" with the unanticipated "after," the problem-solving episode must be past and outcomes amenable to analysis. In a similar way, to appreciate gender or racial bias or other ways that cultural perspectives may sometimes become blindly naturalized in science, one must be at a relatively remote vantage point, to see the culture as culture. History and historical perspective are indispensable for complete nature of science lessons. Everyone here already recognizes this.

But the new assessment format adds clearer focus about the purpose of historical cases and thus how to design them effectively. That is, even though the story is past, the history must be made “present” again. In Latour's language of Janus-faced science, one must
forsake science-made, and restore science-in-the-making. Students must experience a
historically situated perspective, blind to the outcome, akin to the uncertainty in modern cases
that one hopes to inform. For example, to learn about conceptual change, a student ideally
engages in and experiences unforeseen change. Knowing the "right" answer in advance destroys
the lesson, just as a spoiler ruins a mystery or suspense thriller. A case with open-ended
problems, properly contextualized in history, not rationally reconstructed, is essential. Respect
for history is not just the concern of some fussy historians. Rather, as the prototype
assessment remind us, historicism is central to the nature-of-science aims of using history in
science education at all.

Teachers will, nonetheless, need to structure the historical uncertainty mindfully. One
may use successively more complex stages to prepare students for the challenges of modern
decision-making. With intermediate aims, additional assessment strategies may be appropriate.
Most notably, one cannot expect a full appreciation of the nature of science after one case study.
The themes will emerge as a pattern only after many cases, with explicit reflection on each,
where discussion articulates the similarities between them. A first stage of assessment is thus
to consolidate the learning in each case. Here is a sample question, based on the case of
Eijkman and beriberi, familiar to many here(!). It simply lists the NOS features, extracted from
the catalog, and asks the student to articulate them in the particular case. Recall and
understanding: stage 1 and 2 on Bloom’s taxonomy, en route to application, analysis and
evaluation at deeper stages.

As cases accumulate, students ideally draw connections between cases. They apply
NOS knowledge from earlier cases to interpret or solve each new case. By the end of a full-year
course, students should be prepared for a more complete synthesis in a capstone exercise or
final assessment. For example, here is another sample question that I used with my 9th
graders (age 15) in the early 1980s, as part of a final exam: "Discuss the role of controlled
students chose 4 examples from earlier in the class year: including 3 wet labs; 1 dry lab; 2 historical cases; 1 contemporary case. I also used similar questions for other relatively basic concepts: measurement, making connections, factors in scientific belief, and the process of inquiry generally. The students needed to articulate their understanding of the theme and its significance. Assessment of quality here was based on three criteria, identified for the students: first, organization, or unity (does everything support one central idea?); second, breadth, or completeness (were many different ideas related together?); and third, depth, or proficiency (were the ideas conveyed in effective detail?). This style of assessment can empower individual teachers in a local setting. It accommodates them selecting case studies appropriate to their students and that capitalize on their own personal backgrounds and strengths. One need not standardize cases across classrooms. It also ensures that the teacher will be deeply familiar with the examples that a student may use in making comparisons or supporting general statements.

Now, no one should mistake a special focus on NOS as constituting all of science education. Science content and process-of-science skills are equally important. Indeed, students see meaning in nature of science only if it grows out of science. Thus, case studies should integrate all three dimensions — content, process skills, and nature of science — just as one finds them in real scientific practice. In addition, we need to include any and all dimensions of scientific practice that may prove relevant to understanding or assessing its claims. Nature of science is not confined to some short bullet-point list. Nor do we hide errors, unpleasant personalities or unsavory details. Ultimately, we need complete renderings of science — unedited, unabridged, unexpurgated: X-rated science? Not shoehorned into Enlightenment philosophy or biased by myth-conceptions. We need to teach Whole Science. Whole Science, like whole food, includes all essential ingredients. Without supplemental doses of NOS. It
supports healthier understanding. Metaphorically, educators must discourage a diet of highly processed, refined "school science." Whole science as a simply expressed goal also constitutes a powerful argument for including history and nature of science in science education. Namely, do not sell the students short. Feed them Whole Science.

Let me conclude by situating my comments in a concrete context again, by returning to one more prototype question: on the episode dubbed "Climategate." Here, the NOS items include: • human personalities in science, • structure of credibility, • norms of handling scientific data, • robust evidence, • nature of graphs, • credibility of various scientific journals and news media, and • scientific misconduct. Again, knowledge of climate change is not needed. No one needs to evaluate scientific evidence. This case is all about nature of science. But beyond the common consensus list. This case, too, can be informed by history: say, by Millikan’s oil-drop data or Mendel’s data on inheritance in peas; the "tricks" used to map chromosomes or a Mercator-style map; or the publication suppression practices of Newton, Lavoisier or geologist Roderick Murchison.

This example, like the others, reminds us that NOS knowledge is not merely academic. It has real consequences in the real world. In the case of addressing climate change, one might well say it is a matter of life and death. Science education, especially about nature of science, needs to live up to the rhetoric about informing personal and social decision-making. We need to prepare students with episodes of Whole Science from history, rendering fully the details of scientific practice. Yet the authentic context may remind us that one thing is not a part of presenting Whole Science: ideology — whether postmodern cynicism or Enlightenment values.

**Celebrating heroes**, whether it be Galileo, Harvey or Priestley, solely to promote particular visions of science, for example, is an exercise in value-laden history. Neither science nor history, alone, can justify values. Worse, trying to inscribe such personal values into science or history to appear as objective truths betrays the public trust. That is why the proposed
assessment, and the corresponding instruction, focuses on a neutral, well informed analysis, echoing Whole Science.

That, then, is my proposal for rescuing us from the largely unsolved problem of assessment. "Teaching to the test" is often viewed as a liability, taking valuable time away from "teaching." Authentic assessment, however, dissolves this problem. The "test" transparently embodies the goal. Clough and Olson noted in 2008 that "while criticisms of common pencil-and-paper NOS assessments are well placed, attention is needed to creating viable, valid and reliable assessments that will encourage teachers to accurately and consistently implement developmentally appropriate NOS instruction" (p.145). I hope that the prototype and alternatives I have sketched help us move towards that meaningful goal. When our responsibility is met and the test is well written, we will surely endorse, even celebrate, "teaching to the test."

3550 wds [29 mins.]
Table 1. Dimensions of reliability in science.

1. observations and reasoning
   - evidential relevance (empiricism)
   - role of systematic study (versus anecdote)
   - completeness of evidence
   - robustness (agreement among different types of data)
   - role of probability in inference
   - alternative explanations
   - verifiable information versus values

2. methods of investigation
   - controlled experiment (one variable)
   - blind and double-blind studies
   - statistical analysis of error
   - replication and sample size
   - correlation versus causation

3. history and creativity
   - consilience with established evidence
   - role of analogy, interdisciplinary thinking
   - conceptual change
   - error and uncertainty
   - role of imagination and creative syntheses

4. the human context
   - spectrum of motivations for doing science
   - spectrum of human personalities

5. culture
   - role of cultural beliefs (ideology, religion, nationality, etc.)
   - role of gender bias
   - role of racial or class bias

6. social interactions among scientists
   - collaboration and competition among scientists
   - forms of persuasion
   - credibility
   - peer review
   - limits of alternative perspectives and criticism
   - resolving disagreement
   - academic freedom

7. cognitive processes
   - confirmation bias/role of prior beliefs
   - emotional versus evidence-based perceptions of risk

8. economics / funding
   - sources of funding
   - personal conflict of interest

9. instrumentation & experimental practices
   - new instruments and their validation
   - models and model organisms
   - ethics of human subject experimentation

10. communication and transmission of knowledge
    - norms of handling scientific data
    - nature of graphs
    - credibility of various scientific journals and news media
    - fraud or other forms of misconduct
    - social responsibility of scientists
sample assessment question based on single case

What does the case of Christiaan Eijkman & the Cause of Beriberi show about the following aspects of doing biology?

- chance or accident
- theoretical perspectives in interpreting data
- the distinction between causation and correlation
- growth of knowledge through small cumulative additions versus through major conceptual reinterpretations
- the role of individual versus groups in making a discovery
- scientific communication and communities of researchers
- the cultural and economic contexts of science


Discuss the role of controlled experiments using 4 examples:
  
a) you and enzyme catalysis
b) you and the effectiveness of exercise
c) you and spontaneous generation
d) you and nutrient indicators
e) Mendel and inheritance
f) Gause and the outcome of competition
g) Moody and near-death experiences
sample question #2

In November 2009, the author of "The Air Vent," a blog critical of global warming claims, received an anonymous note:

We feel that climate science is, in the current situation, too important to be kept under wraps. We hereby release a random selection of correspondence, code and documents. Hopefully it will give some insight into the science and the people behind it.

The file contained over 1000 e-mails and other material apparently hacked from a server at the Climate Research Unit of the University of East Anglia in Britain. In the e-mails, climatologist Philip Jones, a leading member of the International Panel on Climate Change, included comments about scuttling efforts to release data under a Freedom of Information Act request, a "trick" he used in graphing data, and strategies to limit the publication of critics in peer reviewed journals. James Delingpole, in a blog for England's Telegraph, promptly dubbed it "Climategate." The news sparked a flurry of comments by skeptics who presented this as proof of their repeated claims about fraud, collusion and conspiracy in climate science. Within a week, the term 'Climategate' could be found over 9 million times on the internet.

While sitting at lunch with two co-workers, one mentions how the case just proves that global warming is a joke. The other, an avid environmentalist, contends that scientists don't do things like that, indicating that the posted documents themselves are probably fraudulent. Amid mutual accusations of being misinformed and biased, they ask you set the other straight. Comment on what a well informed interpretation of events in this case might indicate about the conduct of science and the evidence for climate change.

Resource documents include:

- news article from The New York Times
- Delingpole's original blog
- news brief from the journal Nature

See http://ships.umn.edu/knows/