DOCUMENT RESUME

ED 468 034 SE 065 940

AUTHOR Talsma, Valerie L.; Krajcik, Joseph S.

TITLE Comparing Apples and Oranges: Using the National Science Education Standards as a Tool When Assessing Scientific Understandings.

PUB DATE 2002-04-00

NOTE 32p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (New Orleans, LA, April 6-10, 2002). Some figures may not reproduce well.

AVAILABLE FROM For full text: http://www.pitt.edu/~vtalsma/papers/narst2002.pdf.

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS *Academic Achievement; *Academic Standards; *Inquiry; Science Education; Secondary Education; *Student Evaluation

IDENTIFIERS *National Science Education Standards

ABSTRACT

The National Science Education Standards (NRC 1996) present a vision where students are active learners who use inquiry and who create products to represent their emergent understandings. However, educators may find it challenging both to assess student achievement and to communicate students achievement effectively to educational stakeholders. This study employs the Standards as an assessment tool for students' understandings demonstrated in artifacts and as a tool to communicate student achievement. The tool was used to assess and analyze student understandings represented in essays, reports, models, and pre/post instruments across an 18-week watershed investigation. The tool shows some promise in addressing the issues of validity, reliability and impact on instruction. However, tool sensitivity needs further development to distinguish learning below the proficiency level defined by the Standards. Using the Standards as a frame of reference means information about student generated through different assessments in different contexts can now have common meaning and value in the science education community. This contrast with the traditional view of educational measurement that allows for comparisons only when research is based on parallel forms of the same instrument. (Contains 59 references.) (Author/MM)

Comparing Apples and Oranges: Using the National Science Education Standards as a tool when assessing scientific understandings.

Valerie L. Talsma
University of Pittsburgh
Dept. of Instruction & Learning
4C10 Forbes Quad.
Pittsburgh, PA 15260.
Tel: 412-648-7313
Email: vtalsma@pitt.edu

Joseph S. Krajcik
University of Michigan
Educational Studies
610 E. University
Ann Arbor, MI 48109-1259
Ph: 734. 647.0597
email: krajcik@umich.edu

Abstract: The *National Science Education Standards* (NRC 1996) present a vision where students are active learners who use inquiry and who create products to represent their emergent understandings. However, educators may find it challenging both to assess student achievement and to communicate student achievement effectively to educational stakeholders. This study employs the *Standards* as an assessment tool for students' understandings demonstrated in artifacts and as a tool to communicate student achievement. The tool was used to assess and analyze student understandings represented in essays, reports, models and pre/post instruments across an 18-week watershed investigation. The tool shows some promise in addressing the issues of validity, reliability and impact on instruction. However, tool sensitivity needs further development to distinguish learning below the proficiency level defined by the Standards. Using the *Standards* as a frame of reference means information about student achievement generated through different assessments in different contexts can now have common meaning and value in the science education community. This contrasts with the traditional view of educational measurement that allows for comparisons only when research is based on parallel forms of the same instrument.

Keywords: National standards, assessment, student artifacts, conceptual understanding, secondary science

A paper presented at the Annual Meeting for the
National Association for Research in Science Teaching (NARST)
New Orleans, LA - April 2002.

Available online at: http://www.pitt.edu/~vtalsma/papers/NARST2002.pdf

Acknowledgements: We would like to thank the students and teachers of the Traver Creek Project for their participation in this study. This research was supported in part by the Spencer Foundation and the National Science Foundation through the Center for Highly Interactive Computing in Education (hice).

4/12/02

Comparing Apples and Oranges:
Using the National Science Education Standards as a tool when assessing scientific understandings.

Valerie Talsma and Joseph Krajcik

The modes of learning called for in the *National Science Education* Standards (NRC 1996) imply markedly different roles and tasks for students in science classroom and different kinds of student work (Anderson & Helmes, 2001). In this vision, students are active learners who use inquiry to explore authentic problems within a community of scientific practice and who create products (e.g. examinations, journal notes, written reports, diagrams, data sets, physical and mathematical models, and collections of natural objects) to represent their emergent understandings (NRC, 1996). Because of the different roles students are expected to take and the different types of work they are expected to produce, educators may find it challenging both to assess student achievement and to communicate student achievement effectively to educational stakeholders (i.e. colleagues, parents, administrators, researchers, policy makers, etc).

When tasks vary, the assessment of student understandings becomes problematic. In the traditional view of educational measurement, comparisons are allowed only when they are based on parallel forms of the same instrument. Comparing understanding demonstrated in laboratory reports and dynamic computer models is much like "comparing apples and oranges." While the fruits may share some superficial features (e.g. an approximately spherical shape), meaningful comparisons require the applications of more abstract standards, i.e. sugar content, moisture content, or the recommended daily allowance (RDA) of vitamins and minerals.

The students in this study encountered a similar problem. They studied a local creek by conducting a variety of water quality tests, collecting benthic macro-invertebrates and making a series of observations about the physical environment. In this effort, they generated a lot of data, but the numbers and observations did not have meaning beyond a description of the creek. The numbers alone did not tell the students about the water quality. So, the students in this study used "standards of water quality (Mitchell & Stapp, 1994)" in their determination of the health of the creek. These water quality standards allowed the students to make comparisons between specific parameters measured in their creek and scientifically defined values. Students were also able to make comparisons between the different kinds of assessments, e.g. Chemical testing and Bio-Assays, to see if the different forms of assessment led to similar results. By using standards, the students were able to make an assessment of the creek without having to find a comparable creek or without studying values determined before and after an intervention. This idea of comparing observations to defined standards led to the method of assessing student understanding explored in this paper.

Since traditional psychometric techniques cannot be used across dissimilar assessments, comparing student achievement in diverse products may benefit from the application of standards. Standards address what we value as education outcomes and describe how good is good enough (Wiggins, 1993; 1991). Like the water quality standards that students used in their investigations, educational standards provide a frame of reference and a language to compare outcomes across multiple contexts and interventions. The usefulness of the *National Science Education Standards (NRC, 1996)* for describing student achievement across multiple tasks and contexts will depend on how well they meet the measurement criteria of validity, reliability, sensitivity and the impact of the assessment on instruction and classroom practices (See Champagne & Newell, 1992; Haney & Madaus, 1989; Kulm & Malcolm, 1991; Malcolm, 1991; Wiggins, 1993; Talsma and Krajcik, 2002.). If the *Standards* meet these criteria, they may also be useful as a tool to communicate student achievement, providing a language to describe student achievement across different types of student work, time, and contexts. The Standards may also be a tool whereby teachers may report on student progress and achievement to the students themselves, to their colleagues, to parents and to policy makers.

This paper examines the usefulness of the *Standards* for assessing student understandings across multiple artifacts produced in an 18-week investigation of a watershed. This paper is not intended to provide a definitive answer to the usefulness question. Rather, the intention is to provoke discussion and to lay groundwork for using defined standards as a tool in both classroom assessments and educational research on student learning.

Theoretical background

The methods and techniques of measuring learning today represent a movement toward increasing efficiency and making assessments more manageable, standardized, easy to administer, objective, reliable, comparable, and inexpensive (Madaus, 1994). In most school settings, the accepted way for a student to express "understanding" of a history lesson, scientific theory, or novel is to answer questions on a test or perhaps to write an essay (Goldberg, 1992). School assessments usually ask the learner to identify the products (discourse, things, performances) of others; for example, by recognizing the difference between two concepts, by matching scientists with their theories, or by correctly labeling flower parts or vector forces often in an end-of-chapter test (Archibald & Newmann, 1988). In classrooms where the activity of answering recall questions plays a dominant role, this activity often becomes the basis for students' operational definitions of scientific understanding (Anderson & Roth, 1989). Students, who say that they "understand" a concept or topic, often mean that they are prepared to answer recall questions about it; in their experience, this is the sole or primary function of scientific knowledge (Anderson & Roth, 1989).

However, the *Standards* (NRC, 1996) present a different view of scientific understanding, one where students are active learners and creators of knowledge products. In this environment, procedures and situations believed to assess high levels of competence and reasoning abilities, such as artifact assessment, are being re-introduce and advocated by educational researchers and reformers as being more authentic methods of assessment (e.g. Papert, 1991; Perkins, 1992; Wiggins, 1989 & 1993; Wiley & Haretel, 1996).

Critics claim that alternative assessments have too little correspondence to national and state norms, that they can be too subjective and are too inconsistent (e.g. Linn, et al., 1991). Reliability and validity are the key established psychometric criteria for judging the technical adequacy of measures (e.g. see Linn, et al., 1991; Messick, 1989; Messick, 1994). The Burgers (1994) believe that the alternative assessments must be held to the same stringent standards of reliability and validity as those achieved by standardized norm-referenced assessments. In an opposing position, Moss (1994) argues that current conceptions of reliability and validity in educational measurement constrain assessment practices, and these in turn constrain educational opportunities for teachers and students.

Understanding science requires that an individual integrate a complex structure of many types of knowledge, including the ideas of science, relationships between ideas, reasons for these relationships, ways to use the ideas to explain and predict other natural phenomena, and ways to apply them to many events (NRC, 1996). "Scientific understanding" in this study derives from this definition provided in the *Standards* but is informed by the works of many educational researchers (e.g. Schwab, 1964, Schoenfeld, 1985, Posner, et al., 1982; Brown, Collins, and Duguid, 1989; White and Gunstone, 1992; Perkins & Simmons, 1988, Perkins, et al. 1995; Eisenhart, et al. 1993; Novak & Gowin, 1984). For the purpose of this study, scientific understanding was defined as the set of elements a learner possesses about a concept and the richness of interconnections and relationships made between concepts. Implicit in this definition is the idea that understandings are dynamic rather than static, for new knowledge can be added to the set, new links can be formed between things already known, and the knowledge set can be restructured based on more abstract principles.

One way to better understand "understanding," is to contrast this construct with two other constructs, "knowing" and "remembering." To know or to remember something suggests that one has information in storage, such as a phone number or an author and book title, and can retrieve it on call (Perkins, 1991). Scientific knowledge refers to facts, concepts, principles, laws, theories, and models (NRC, 1996). A learner, who *knows* and can *remember* scientific knowledge, can recite it (e.g. Avogadro's number is 6.02 x10[e]23 or pH is measured on a scale of 1-14). A learner who *understands* the scientific knowledge can use that knowledge to do something effective, transformative, or novel with a problem or complex situation (e.g. use a pH measurements of a creek system to predict which macro-invertebrates may be found there) (Wiggins, 1989).

Understanding goes beyond knowing or retrieving information along a continuum, which includes readiness for a wider range of performances (Perkins, 1991). For example, suppose that a learner can *explain* a concept (e.g. dissolved oxygen) in their own words (not just reciting a canned definition), can *exemplify* its use in fresh contexts (aquariums instead of streams), can *make analogies* to novel situations (carbonated beverages, stuffy rooms), can *generalize* the law (solubility of gasses), recognizing other laws or principles with the same form (solutions), most educators would agree that learner has an understanding of the construct in question. Understandings can be demonstrated because understanding involves action more than the possession or accumulation of cued knowledge (Perkins, 1991; Wiggins, 1993).

The *Standards* claim that inferences about students' understandings "can be based on the analysis of performance in the science classroom and work products" (Ch 5). A few studies (Spitulnik, 1995; Spitulnik, et. al., 1996; Stratford, 1996) have examined student understandings exhibited in discrete artifacts (dynamic models, hypermedia documents, etc.) but not across a series of artifacts that represent understandings across a longer time period.

This paper explores using an assessment method where the *Standards* are employed as a tool to measure students' scientific understandings.

Study Design

This paper rose from an investigation of the breadth and depth of scientific understandings acquired by high students engaged in extended inquiry around a creek (Talsma, 2002). The creek project was chosen for study because (1). the content was interdisciplinary, combining content from earth science, biology and chemistry; and (2) the project had the potential of meeting a number of science standards. The guiding questions for the study were:

How well do the standards capture the content of the creek curriculum?

What scientific understandings, breadth and depth, did students demonstrate in the artifacts?

How well did these understandings map on to the Standards*?*

These questions helped to frame the data collection and analysis. Data were collected in four ninth grade classrooms ($n_{students} = 99$) in one school enacting a project-based science (Blumenfeld, et al., 1991; Huebel-Drake, et al., 1995; Krajcik, et al., 1998; Marx, et al., 1997) study of a watershed over the course of one semester. Multiple sources of qualitative and quantitative data were collected, including: student constructed artifacts - essays, scientific reports, and computer models; pre- and post-instruments; classroom observations and classroom handouts (Talsma, 2002).

The guiding questions helped to frame a four-step analysis process of the data: (1) The delineation of project curriculum and mapping it onto the *Standards*, (2) The identification of opportunities (and expectations) to demonstrate understanding in the selected artifacts and a pre/post test instrument. (3) Analysis of student understandings in each of the artifacts scored on a four-point scale. And (4) the examination of student understandings across time and artifacts. Each of the five steps and the resulting findings are addressed individually in the next sections. At each step of the analysis, problems about using the *Standards* as an assessment tool were encountered. These are discussed in the context in which they arose and the solutions that were employed in this study are described.

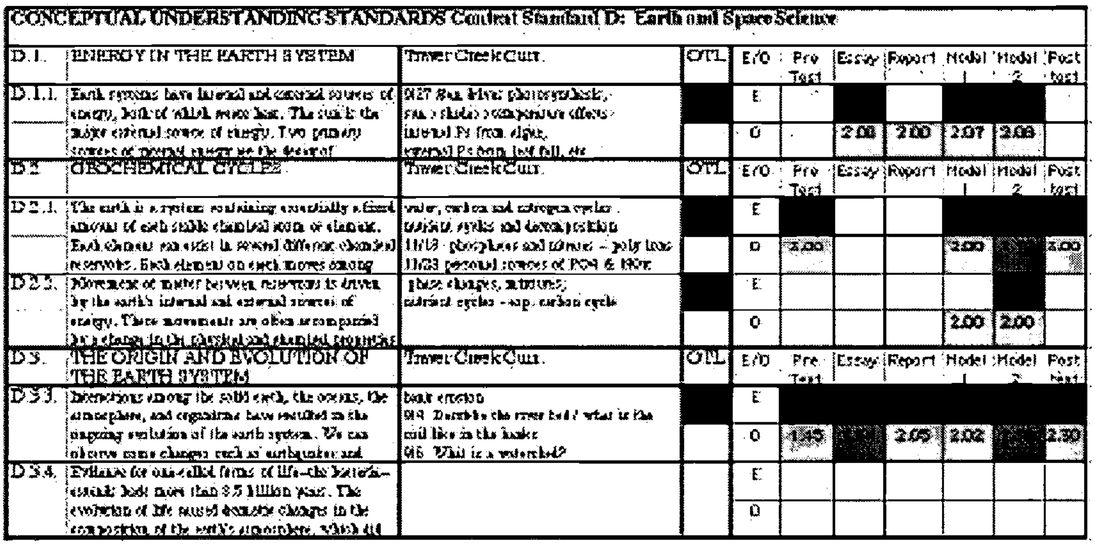
Because students worked individually, in pairs, or in groups of 4-5 students on the different assessment measures, three abbreviations are used in the data: n_a represents the number of artifacts in the analysis, n_s represents the number of students in the analysis, and

n_c represents the number of paired cases (student demonstrating a *Standards* understanding on two different artifacts).

(1) The delineation of project curriculum and mapping it onto the *Standards*

The classroom observations, videotapes, and collections of student handouts and teacher's notes were used to characterize and map out the content of the Creek Project. The project content was then compared to the set of outcomes that students should know, understand, and be able to do in natural science in grades 9 through 12 in the *National Science Education Standards* (Chapter 6, NRC, 1996). Applicable standards were identified and used to create a data matrix of conceptual understandings. A small section of the *Standards* and curriculum matrix is shown in Table 1.

Table 1: Part of the matrix used to map curriculum onto the *National Science Education Standards* (NRC, 1996). This part comes from "Content Standard D: Earth and Space Science." Individual Standards are represented by a three-character code. The code for each standard was not given in the official document but was derived by using the *Standards'* major designations (Contend Standards A-G) and then sequentially numbering the sub-standards below each designation.



CONCEPTUAL UNDERSTANDING STANDARDS Content Standard D: Earth and Space Science										
D.1.	ENERGY IN THE EARTH SYSTEM	Traver Creek Curr.	OTL	E/O	Pre Test	Essay	Report	Model 1	Model 2	Post test
D.1.1.	Earth systems have internal and external sources of energy, both of which create heat. The sun is the major external source of energy. Two primary sources of internal energy are the decay of	9/27 Sun. Rivers: photosynthesis, sun > shade > temperature effects > internal Es from algae, external Es from leaf fall, etc	■	E		■		■	■	
				O		2.00	2.00	2.07	2.08	
D.2	GEOCHEMICAL CYCLES	Traver Creek Curr.	OTL	E/O	Pre Test	Essay	Report	Model 1	Model 2	Post test
D.2.1.	The earth is a system containing essentially a fixed amount of each stable chemical atom or element. Each element can exist in several different chemical reservoirs. Each element on earth moves among	water, carbon and nitrogen cycles. nutrient cycles and decomposition 11/13 - phosphates and nitrates - poly [illegible] 11/23 personal sources of PO4 & NO3	■	E	■			■	■	■
				O	2.00			2.00	▨	2.00
D.2.2.	Movement of matter between reservoirs is driven by the earth's internal and external sources of energy. These movements are often accompanied by a change in the physical and chemical properties	phase changes, mixtures; nutrient cycles - esp. carbon cycle	■	E					■	
				O				2.00	2.00	
D.3.	THE ORIGIN AND EVOLUTION OF THE EARTH SYSTEM	Traver Creek Curr.	OTL	E/O	Pre Test	Essay	Report	Model 1	Model 2	Post test
D.3.3.	Interactions among the solid earth, the oceans, the atmosphere, and organisms have resulted in the ongoing evolution of the earth system. We can observe some changes such as earthquakes and	bank erosion 9/9 Describe the creek bed / what is the soil like in the banks 9/6 What is a watershed?	■	E	■	■	■	■	■	■
				O	1.45	▨	2.05	2.02	▨	2.30
D.3.4.	Evidence for one-celled forms of life—the bacteria—extends back more than 3.5 billion years. The evolution of life caused dramatic changes in the composition of the earth's atmosphere, which did			E						
				O						

In the mapping for the course curriculum onto the *Standards,* the first problem in using this tool was encountered. This problem was the lack of correspondence between content in the *Standards* and scientific content found in the curriculum. For example, one of the first inquiry activities the students did was watershed mapping using topographic maps. Students identified the boundaries of the creek's catchment basin, measured the catchment area and stream length, calculated stream gradient and percent of major land uses, and determined stream order at major road crossings (class handout and field notes 9/6/96).

The *National Science Education Standards* (NRC, 1996) outline 126 science standards for grades 9-12. But no where is the concept of "watershed" found in the standards. The closest match is Standard D.3.3 – Interactions of earth systems (Table 1). However, this standard address more the process that first produce a watershed and then account for changes in the watershed. The solution to correspondence problem was to use interpretations of the *Standards* to map course content where a fit could be found. For example, landuses in the watershed mapped onto standard F.3.3. - Humans use many natural systems as resources. Otherwise, the content, like the concept of "watershed" was noted as falling outside the standards.

When aligned with the *Standards,* the creek curriculum was found to address seventy-five (75) standards at the high school level, about 60% of the content explicated by the standards as being important for students to understand. These were distributed across the seven major content standards as summarized in Table 2.

Table 2: The Creek Project curriculum addresses National Science Education Standards (NRC, 1996) across each of the major content divisions.

Major NRC Standards addressed by the Creek Project	Number of 9-12 Standards	Number of standards in the curriculum
Standard A: Science as Inquiry	12	10
Standard B: Physical Science	28	11
Standard C: Life Science	28	19
Standard D: Earth and Space Science	13	4
Standard E: Science and Technology	10	10
Standard F: Science in Personal and Social Perspectives	25	14
Standard G: History and Nature of Science	10	7
Total	126	75

It is interesting to note that although the creek curriculum content was relatively balanced between earth science, biology, and chemistry in terms of classroom time and number of concepts addressed, this balance was not evident when looking at the number of standards addressed. When mapped onto the *Standards*, the creek curriculum appears heavily weighted toward the Life Sciences (19 standards in Standard C) and Environmental Sciences (14 standards in Standard F) (Table 2). This was due in part to how well the content has been delineated among the different substandards in each section. For example, in the Life Sciences, Standard C.4 addresses the interdependence of organisms and there are five objectives that differentiate the various kinds of interactions. All five substandards match content addressed in the creek project. At the same time in Standard D on Earth and Space Science, weather phenomena, which were addressed in the context of the impacts a watershed due to flooding, drought, etc, were only addressed tangentially in three standards (D.1.1, D.2.1 and D.3.3). In addition, over three weeks of classroom time were devoted to introductory chemistry (atoms, compounds, simple reactions, etc.) yet only two of the five sub-standards under B.3 Chemical Reactions applied to this project.

The standards that were assessed in the four artifacts and the pre/post tests formed a subset of the standards addressed by the project. These measures are examined next.

(2) The identification of opportunities to demonstrate understanding in the selected artifacts and a pre/post test instrument.

Once the content of the curriculum was mapped out, the different assessments were examined for the opportunities and constraints provided by the task structure for students to express their scientific understandings. The analysis of the artifact affordances included close examination of written materials such as project guidelines and handouts, assessment criteria presented by the teachers, teacher explanations in class, and the capabilities and features of the technology employed in the artifact construction (e.g. ClarisWorks® spreadsheets and charts for the stream reports and features of Model-It® for the creation of the computer models.)

For example, in the report assignment, students were given a detailed handout describing the parts of their reports. This handout was reviewed during class time and additional examples were given. The details include very specific instructions and check lists. For example, for the report introduction, students were directed as follows:

> INTRODUCTION (About 2 paragraphs)
>
> The introduction should provide a context for the topic under study. The introduction provides the background necessary to understand the rest of the report. In addition the introduction should provide a concise <u>statement of the problem</u>. That is, tell precisely what questions you are trying to answer. Suggestions for what you should have in the introduction:
>
> ____What question were you trying to answer about Traver Creek?
>
> ____ A description of what benthic organisms are and why they are used as indicators of health for the creek.
>
> ____A description of physical forces that influence the benthic community.
>
> (Class handout 10/21/96.)

In this piece of the assignment, it is possible to identify multiple matches to the standards. The requirement to provide a concise statement of the problem or question maps onto standard A.1.1 "Identify questions and concepts that guide scientific investigations." The requirement to provide a context or purpose for the study maps onto standard A.2.2. "Scientists conduct investigations for a wide variety of reasons." The descriptions of benthic organisms maps onto standard C.3.5 (biological classifications). Why benthics are used as indicators of a healthy creek maps onto standard C.5.5 (niche concept - limiting factors). The description of the physical forces that influence the benthic community maps onto standards D.3.3 (interactions of earth systems) and C.5.5 (niche concept - limiting factors). If students followed these guidelines, they could potentially demonstrate deeper understandings of biological classification[1] (Standard C.3.5). These directions are very specific and the content is relatively easy to map onto the standards.

Twice during the semester, students created dynamic computer models. These computer models were constructed using Model-It 3.0b software which provided a dynamic modeling environment designed specifically for learners who are unfamiliar with dynamic modeling and lack mathematical or symbol manipulation skills (Jackson, Krajcik & Soloway, 1998). Models consisted of objects ("things" in the system being modeled such as the creek, fish, people), factors (measurable attributes of objects, e.g. creek temperature and pH, number of fish), and relationships between factors (e.g. as water temperature increases, dissolved oxygen decreases).

Both modeling assignments provided students with opportunities to demonstrate understandings about the processes of science and the nature of science. Understandings that map onto Standard E.1 (Abilities of technological design) were most readily identified in the assignment and the affordances provided by the Model-It software (Table 3). In the planning notebook of the model, students were expected to describe a goal for their model (Standard E.1.1). They were then expected to begin planning their model by identifying important objects and factors (Standard E.1.2). By constructing the model, they implemented a proposed solution (Standard E.1.3). Finally, they were asked to thoughtfully evaluate their model (Standard E.1.4).

In the students' goal statement in the model's notepad (Figure 1), they needed to describe a purpose for their model (Standard A.2.2). Periodically during model construction, students were reminded by both the software and the teachers to test their models. Testing and revising the model to better reflect their understanding would reveal a performance understanding about students' criteria for scientific explanations (Standard G.2.2). In their evaluation of their model they were asked how they would change their model. Statements about change would reflect students' understandings about the tentative nature of scientific knowledge (G.2.3).

However, for the two modeling assignments, the directions for content were much less specific than the report assignment. During the first model building cycle (week 11), students were given one long (90 min.) and one short (45 min.) class period to build a model of the physical and biological factors in the study creek and to test their models. The handout for this assignment specified that:

> *Your model must convincingly demonstrate your understanding of the physical factors you've chosen to model and how they relate to the benthics. Make sure that you fill out a plan, describe each object and factor, explain*

[1] BMI's include orders of insects (mayflies, stoneflies, odonates, hemipterans, dipteans, etc.), crustaceans (isopods, amphipods, crayfish), mollusces (snails, clams, limpets), and annelids (tubifex, leeches).

> *each relationship, test your model as you go along, and evaluate it at the end. (Emphasis in original - Class Handout 11/4/1996)*

A model that would evidence the required conceptual understandings would include one or more physical factors and show a relationship between that factor and the benthics (object). The interactions of physical and biological components of the system would map onto standard D.3.3 (interactions in earth systems) and C.5.5 (niche concept - limiting factors) (Table 3). If students included physical factors such as the sun affecting the temperature of the water, their understandings might also map onto other earth science standards (i.e. D.1.1- sources of energy and D.2.1- conservation of matter). If they elaborated on the benthic macroinvertebrates, they might also represent other Life Science understandings such as those under C.4 - The interdependence of organisms.

The second modeling assignment at the end of the semester was even broader. For this assignment, students were asked to build a model that demonstrated in-depth understanding of a stream ecosystem and that included physical, chemical, and biological factors of the stream. This modeling assignment was much more open in terms of which conceptual understandings students might include. As such, students might address a number of different conceptual understandings that would map unto physical science (Standard B), Life science (Standard C), Earth Science (Standard D), and Environmental science (Standard F) (Table 3). However, there were a few standards, such as B.1.1 on the structure of atoms, B.2.1 on chemical reactions and B.2.2 on the Periodic Table that included content that could not be reasonably represented within the modeling environment. These standards are examples of those few that were not supported by this assignment. Although students were offered the opportunity to include content from 26 different standards in their models, it was not expected for them to actually do so. Rather, the assignment was designed for students to select content from the biological, physical and chemical factors of stream phenomena, representing a subset of content standards in any one model.

Herein lies an issue in artifact assessments, how to deal with the lack of specificity in requirements, the "ifs" and "mights"? Some content was specified or required by a task, like the descriptions of the benthic organisms in the report assignment. Some content might be expected in an artifact, like the sun's effect on the creek in the first modeling assignment. Other content may opportunistically arise as learners refine, extend and elaborate their artifacts. And there were content standards that could not be represented in a particular artifact because of constraints in the media or the task structure.

The required elements were easily handled. Rubrics were created and demonstrations of understanding checked off. But keeping track of serendipitous pieces of scientific understanding was more problematic. When did a statement represent a unique element of understanding and when should it be grouped with other statements of understandings? The *Standards*, at least, provided a finite set of expected understandings onto which different representations could be mapped.

Table 3: ***National Science Education Standards*** **(NRC, 1996) from the creek curriculum assessed in the artifacts and pre/post test assessments.**

NRC standards assessed in the Creek Project	Essay	Report	Model 1	Model 2	Pre/ Post
Standard A: Science as Inquiry	A.1.1	A.1.1			A.1.1
		A.1.2			A.1.2
		A.1.3			
	A.1.4	A.1.4			A.1.4
	A.2.2	A.2.2	A.2.2	A.2.2	A.2.2
		A.2.5			A.2.5
Standard B: Physical Science					B.1.1
					B.2.2
				B.2.5	B.2.5
					B.3.1
				B.3.3	
Standard C: Life Science	C.3.5	C.3.5		C.3.5	C.3.5
				C.4.1	
		C.4.2	C.4.2	C.4.2	
			C.4.3	C.4.3	C.4.3
				C.4.4	
	C.4.5		C.4.5	C.4.5	C.4.5
				C.5.2	C.5.2
					C.5.4
				C.5.1	C.5.1
		C.5.5	C.5.5	C.5.5	C.5.5
				C.5.6	C.5.6
Standard D: Earth and Space Science	D.1.1		D.1.1	D.1.1	
			D.2.1	D.2.1	D.2.1
				D.2.2	
	D.3.3	D.3.3	D.3.3	D.3.3	D.3.3
Standard E: Science and Technology			E.1.1	E.1.1	
			E.1.2	E.1.2	E.1.2
		E.1.3	E.1.3	E.1.3	
			E.1.4	E.1.4	
Standard F: Science in Personal and Social Perspectives	F.1.3				
				F.2.1	F.2.1
	F.3.1			F.3.1	
	F.3.3			F.3.3	
		F.5.2		F.5.2	F.5.2
		F.5.3			F.5.3
					F.6.1
					F.6.4
	F.6.5			F.6.5	
Standard G: History and Nature of Science			G.2.2	G.2.2	
			G.2.3	G.2.3	
Totals Standards in Assessment	11	13	14	28	25

(artifacts =40; pre/post = 25)

This study chose to identify all required elements and all plausible representations of content as "Expected" content in the artifacts. By this classification, the *Standards* that were assessed in the four artifacts and the pre/post tests formed a subset of the seventy-five *Standards* addressed by the project. All together, the four major artifacts provided opportunities for students to demonstrate understandings on 40 standards while the pre/post only assessed understandings on 25 standards (Table 3).

All four artifacts assessed only one common content standard (D.3.3 - Interactions of earth systems) and one nature of science standard (A.2.2 – Understandings about scientific inquiry). Five standards (C.3.5, C.4.2, C.4.5, C.5.5, D.1.1) were assessed by three of the four artifacts and seven (C.4.3, D.2.1, F.3.1, F.3.3, F.5.2, F.5.3, F.6.5) were assessed by two of the four artifacts (Table 3). The remaining standards assessed in the artifacts were supported by a single artifact, most often in second modeling assignment. From this distribution of standards, there does not appear to be an over-representation of specific content across the different artifacts (Standard D.3.3 on the interactions of earth system incorporates a broad array of potential conceptual understandings). There does appear to be an under-representation of project content in the artifacts, especially representation of content in Standards B and F.

The content in the Physical Sciences (Standard B) represent several weeks of classroom instruction (November 13- December 6) so the lack of artifacts that would assess students' conceptual understandings in this area is a potential weakness of this approach. Students did complete some smaller artifacts during this period including quizzes and mini-lab write-ups that provided the classroom teachers with some assessment of students' chemical knowledge.

(3) Analysis of students' conceptual understandings in each of the artifacts.

Students' conceptual understandings were derived from content represented in their artifacts, especially student descriptions and explanations of phenomena. The artifacts were carefully examined and content was mapped onto the standards. Understandings were identified when two or more ideas were connected. Thus, lists of observations that more closely resemble a note taking assignment were not counted as understandings. But when students made connections between two or more ideas, like shade affecting the level of dissolved oxygen in the creek, these were considered evidence of understanding and mapped onto the appropriate standards (e.g. B.2.5 and D.3.3).

After all the standards were identified in an artifact, the quality of understanding for each standard was determined. In this effort, another issue of using the *Standards* as a tool arose. Although the *Standards* claim to be criteria by which to judge the quality of what students know and are able to do (NRC, 1996, Ch 5) they are, in fact, content standards and not performance standards. "Content standards" specify "what" students should know and be able to do (NESIC, 1993). They indicate the knowledge and skills -- the ways of thinking, working, communicating, reasoning, and investigating, and the most important and enduring ideas, concepts, issues, dilemmas, and knowledge essential to the discipline -- that should be taught and learned in school (NESIC, 1993, p. ii). As written, the *Standards* are essentially a threshold. Either students are achieving at the level of the standards or they are not.

"Performance standards," in contrast to content standards, specify "how good is good enough" (Shavelson, Baxter, & Pine, 1992; Wiggins, 1991). They provide not only models but also a set of implicit criteria against which to measure achievement (Wiggins, 1991). Performance standards relate to issues of assessment that gauge the degree to which content standards have been attained, the indices of quality that specify how adept or competent a student demonstration must be. A performance standard indicates both the nature of the evidence (such as an essay, mathematical proof, scientific experiment, project, exam or combination of these) required to demonstrate that the content standard has been met and the *quality of student performance* that will be deemed acceptable (NESIC, 1993, p. iii). Progress involves successive approximations in the direction of an exemplary performance (Wiggins, 1991).

Thus, in order to classify the quality of understandings demonstrated in the student artifacts, the *Standards* had to be changed from content standards to performance standards. A four level (0-3) coding scheme, adapted from Stratford (1996), Carey, et al (1989), Grosslight, et. al, (1991) and Spitulnik (1998), was used to classify the "quality" of the understandings. These levels included:

Level 3: Representation is scientifically correct to the level used in the *National Science Education Standards* (NRC, 1996) and contains no extraneous or incorrect ideas, statements concur with expert propositions (proficient or mastery level).[2]

Level 2: Representation is partially correct but is missing critical information OR contains some extraneous and/or incorrect information (developing).

Level 1: Representation contains substantial errors OR fundamental differences between the students' and expert's conceptions as depicted in the *Standards* (non-scientific or novice).

Level 0: Student did not provide a representation OR if some representation is given, it does not evidence understanding, perhaps nonsensical, e.g. "Ugh" (no evidence). Level 0 does NOT mean the student does not understand the content.

[2] Note that a Level 3 code does not represent the highest quality of understanding . A content specialist would naturally demonstrate higher levels of understanding. The level 3 represents a threshold value for the quality of scientific understanding expected of high school graduates by the *Standards*. In addition, a Level 0 understanding simply means that specific content was not represented in an artifact. It does not mean that a student doesn't understand the content.

For example, if a student wrote, "We found high levels of dissolved oxygen in our section of the creek. We think this might be because we had a lot of shade in our section." This statement would be scored as Level 2 for B.2.5 dealing with gas solubility because there is no explanation of causality and Level 2 for standard D.3.3 for identifying an interaction between earth systems (biotic shade influencing abiotic DO_2.). If instead the students had written:

> *We found high levels of dissolved oxygen in our section of the creek. We think this might be because we had a lot of shade in our section. Trees create shade which blocks the sun from reaching the creek. Because the water is shaded from the sun, it doesn't warm up as much. And cooler waters can hold more dissolved oxygen than warm waters.*

They would reveal a more robust understanding of the relationship between trees, shade, temperature, and dissolved oxygen. The second example would map onto standard D.1.1 - Sun as source of energy - Level 3, D.3.3 - interaction of earth systems - Level 3 (shade influences water temperature and thus DO levels), and B.2.5 - gas solubility - Level 2. A standard B.2.5 Level 3 would be recorded if a student had correctly explained why warm waters hold less dissolved gasses than cool waters.

The general four level scale worked well for the content standards (B, C, D & F). More specific rubrics were established for some of the process and nature of science standards (Table 4). For example, Standard A which includes understandings and abilities to do science (A.1) and understandings about scientific inquiry (A.2), was assessed in the four artifacts (Table 3). Each sub-standard was given a four level performance rubric based on models in the literature. For instance, Carey, et al. (1989) explored 7th grade students' understanding about the nature of scientific knowledge and inquiry. They described three levels of understanding about scientific experimentation that students might exhibit. These levels were used to construct the performance levels of Standards A.2.2 and A.2.5. shown in Table 4.

A second educational researcher provided inter-rater reliability on the artifact scoring. After instruction in the basics of stream related scientific understandings, an introduction to the scoring guide, and a few practice artifacts, the researcher and first author independently scored a subset (~10%) of the artifacts. Pearson product-moment correlation coefficients were calculated for agreement on each standard scored in a particular measure. Inter-rater reliability on conceptual understanding standards covered a fairly high range ($0.837 \leq r \leq 0.958$ or $.70 \leq r^2 \leq .92$).

Table 4: Performance standards for Standards A, E and G.

NRC Standard Codes	Level 3	Level 2	Level 1	Level 0
Problem Definition A.1.1 E.1.1	defines a reasonable, well focused problem area to be addressed within constraints, identifies essential elements of the problem makes a prediction, stating possible outcomes,	defines , through revision, a reasonable problem area to be addressed within constraints identifies essential elements of the problem states reasonable thesis, no prediction, may be general	has difficulty defining a reasonable problem area given constraints, problem may be too broad vague, general, undefined or non-scientific thesis,	non-existent, no evidence does not define a problem
Planning: designs a method or approach A.1.2a E.1.2	designs and implements a method to address problem, including gathering resources, synthesizing information, organizing and presenting findings. suggests an experimental design that directly addresses identified problem suggests data . Explanation integrates and applies knowledge, controls variables.	designs a method to address problem, may have trouble getting started Reports on method used employs some methodology, perhaps survey, - suggests data to collected. may mix scientific and non-scientific approaches. may not control variables	suggests nonscientific approach, eg. reading, talking to people had difficulty designing a method to address problem methods section vague/general	non-existent, no evidence does not attempt to address problem lacks a coherent design
construction of an explanation, argument (A.1.4) or model (E.1.3).	supports with empirical evidence, uses empirical evidence and/or models to justify or evaluate an argument or stated position constructs a model with explanatory power, elegance and parsimony	uses some evidence to justify or evaluate an argument or stated position constructs a model with some detail and explanatory power; includes too much detail so that explanatory power is lost.	does not state a position does not use evidence to support an argument or position constructs a simple model with little detail and no explanations	non-existent, no evidence

Table 4con : Performance standards for Standards A, E and G.

Summarizing and/or conclusions A.1.5 E.1.4 (models)	making a conclusion about and explaining an everyday situation by extending generalizations constructed in models, explanations Supported - predicts results based upon design that would support or refute hypothesis	making a conclusion about and explaining an everyday situation using some supporting evidence	conclusion is not consistent with evidence or prior arguments.	no conclusion, no evidence.
describes purpose A.2.2	experiments as hypothesis testing or exploration model is constructed in the service of developing and testing **ideas**	experiments test an idea to see if it is right (verification) specific, explicit purpose for model but focus on reality, not ideas	purpose for experiment is to do the experiment identifies no purpose beyond class expectations Models are simple copies of reality	non-existent, no evidence
explains rationales A.2.5	evaluating which of several designs could be used to serve the purpose ("to see.., so that", etc.) uses empirical evidence to justify or evaluate a design or stated position	uses some evidence to justify or evaluate a design or stated position	does not use evidence, or does not justify or evaluate a design or stated position	non-existent, no evidence

Two of the major student artifacts constructed during the creek curriculum were computer-based models of stream phenomena. A relatively simple model created by Chase[3] illustrates how student understandings were identified and characterized in their models.

Chase's Model of the Effects of Forest Fires

For his final model, Chase, a male student working alone, decided that he wanted to create of model "to show how a forest fire would affect various characteristics of the creek." In the planning of his model, Chase defined his purpose/ problem and he began to plan the model by filling out the fields in the planning window of the Model-It software (Figure 1). The scientific understandings demonstrated in this part of the model include:

Problem Definition - because Chase required some discussion with the classroom teacher about his problem (Field notes 1/8/97) and because he did not include a prediction of how he expected the fire to affect the

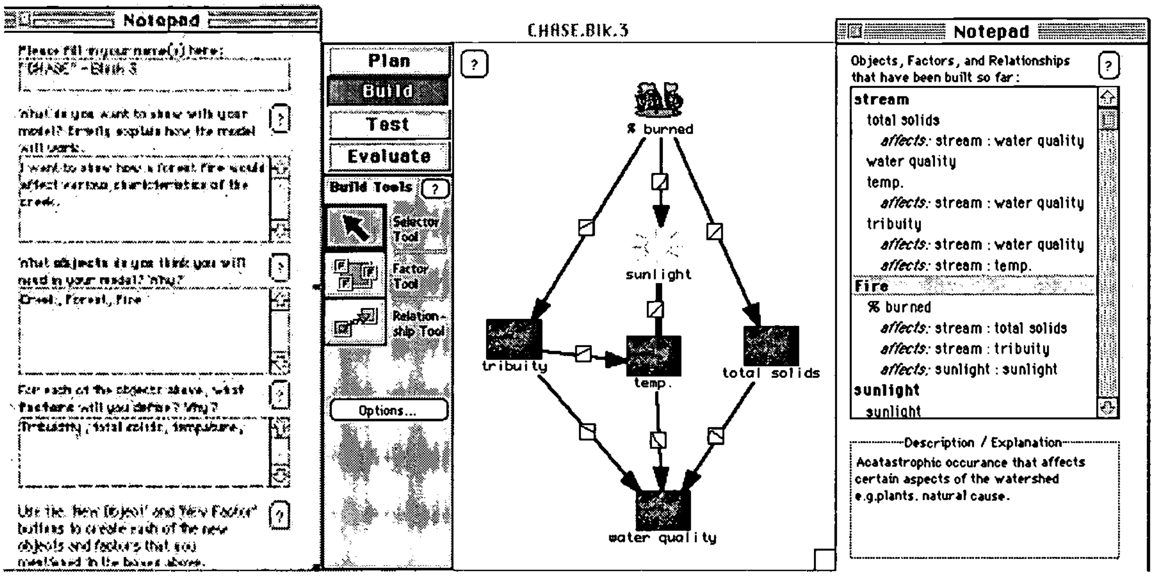
[3] All names are pseudonyms.

creek (Figure 1), his understanding of standard E.1.1 (Table 4) was coded at level 2.

Purpose of the model - In choosing to model a forest fire, Chase was pursuing a hypothetical situation. Therefore, his model was interpreted as "developing and testing ideas (theories, possibilities)" - Level 3 understanding as opposed to a focus on reality (Level 2) for standard A.2.2. (Table 4)/

Planing the model (Objects & Factors). In the third and fourth fields of the plan notepad Chase identified 3 objects and 3 factors for his model (Figure 1). These included two of the three objects and 3 of the 6 factors he actually included in his model. Thus, his understanding on standard E.1.2 was coded at Level 2 - 33-66% of objects and factors identified in plan (Table 4).

In the Model, Chase created three objects: stream, using a digital image from the class sever; fire - modifying a ClarisWorks clip art; and sunlight - clipart. Chase did not include any rationales for any of the objects and factors in his plan (Answers to the "Why?" question in the prompt). Therefore, his demonstrated understandings for standard A.2.5 was coded at Level 0 for no evidence (table 4)



Notepad

Please fill in your name(s) here:
"CHASE" - Blk 3

What do you want to show with your model? Briefly explain how the model will work.
I want to show how a forest fire would affect various characteristics of the creek.

What objects do you think you will need in your model? Why?
Creek, Forest, Fire

For each of the objects above, what factors will you define? Why?
Tributaty, total solids, temprature,

Use the "New Object" and "New Factor" buttons to create each of the new objects and factors that you mentioned in the boxes above.

Notepad

Objects, Factors, and Relationships that have been built so far:

stream
total solids
affects: stream : water quality
water quality
temp.
affects: stream : water quality
tribuity
affects: stream : water quality
affects: stream : temp.
Fire
% burned
affects: stream : total solids
affects: stream : tribuity
affects; sunlight : sunlight
sunlight
sunlight

Description / Explanation
Acatastrophic occurance that affects certain aspects of the watershed e.g.plants. natural cause.

Figure 1: Chase's Model with planning notebook on the left.

Chase require a little assistance in manipulating the software during the beginning of the building phase until he regain familiarity (Field notes 1/11/97). He then proceeded to construct a model with three objects, six factors, and nine relationships (Figure 1). In the descriptions of the factors and relationships, Chase demonstrated his scientific understandings. For example, in his creation of the factor, "stream: total solids", Chase demonstrated his understanding in two ways: in his description of the factor, "Total solids are the dissolved matireals[sic] and the suspended matireals[sic]" and in his definition of the range, quantitatively from 0 to 500 mg/L. Since both Chase's description and his defined range were scientifically accurate, and because his understanding of total solids maps onto the Standard on states of matter/ mixtures and solutions (B.2.5), Chase's understanding on this standard was coded at Level 3.

Chase also demonstrated his understandings in the relationships between factors. For example, in the relationship between total solids and water quality, Chase demonstrated a high degree of understanding (Level 3) in three areas: he created a scientifically accurate relationship (Mitchell & Stapp 1994, p. 84.); he provided an accurate explanation for this relationship and he provided an elaborate explanation by listingmultiple causes: the reduction of water clarity due to increasing turbidity, a decrease in photosynthesis rates caused by a reduction in sunlight penetration, the possibility that these materials will bind to pollutants, and an increase of stream temperature caused by the absorption of sun energy of the total solids. The content of this relationship maps onto Standard D.3.3 - Interactions of Earth Systems.

In the earth and space sciences, Chase's model addressed Standard D.1.1 - Sources of Energy. This understanding was demonstrated in three relationships: stream: tribuity [sic - turbidity] affects stream: temp[ature], Fire:%burned affects sunlight:sunlight, and sunlight:sunlight affects stream:temp (Figure 1). In these three relationships, while the overall sense of the relationships is correct, there are some small errors. For example in the relationship between Fire:%burned and sunlight:sunlight, the shape of the relationship should be "increases by a little" not by "more and more." The banks of the stream comprise only a small part of the watershed that is burned, but this is the only area that would be shading the steam and thus affecting sunlight so although a large amount of the watershed might burn, a much smaller percent of that affects the sunlight reaching the stream. Another relationship, between sunlight and temperature, lacked an explanation. For these reasons, Chase's demonstrated understandings along Standard D.1.1 were coded at Level 2.

The Model-It 3.0b environment also afforded Chase the opportunity to test his model. On day two of model construction, Chase conducted a test of the relationships emanating from the factor "fire:%burned." Chase selected relationships, opened meters, started the test, changed "fire:%burned" from 0 to 49%, stopped the test and then created a new relationship between turbidity and stream temperature. Later, Chase tested this new relationship. He then created the factor "stream:water quality" and built the three relationships that affect it. He ran a final test of the model after which he edited the factors "fire:%burned" and "stream:water quality." Because Chase tested his model, it was possible to make a determination of his demonstrated understanding about the nature of knowledge in models (standard G.2.2). In the sequence described above, Chase tested his model and revised it to produce a desired outcome. There was no evidence that Chase was testing ideas and revised his model to better account for evidence (Level 3 - Table 4). Therefore his demonstrated understanding for standard G.2.2 was coded at Level 2.

In his model evaluation Chase responded to two questions. In the first, "How well does your model work, or if it doesn't, why not?" Chase replied, "It worked well I enjoyed building it and it showed me how a creek is affected by fire." This response was classified as a Level 2 understanding for strategic understanding standard E.1.4 because, although Chase did compare the model to his purpose, he did not provide any evidence for how well the model worked. In the second question, "What would you change to make your model work better or be more complex?" Chase wrote, "I would make the amount of sunlight start a little higher seeing as a real creek is not totally blocked from the sun by trees." This response was coded at Level 2 for the standard G.2.3 on the tentative nature of science because this would not involve a substantial change to the model and he does not indicate that he would use empirical values in the model (Level 3 understandings – Table 4).

Chase's model included factors and content from the physical and chemical assessment of the creek. His model did not explicitly include biological factors, which were part of the assignment. Never the less, Chase's overall model showed an elegance and parsimony that explained the essential effects of a forest fire on the creek. Therefore, for standard E.1.3 - Implementing a plan/building the model, Chase's demonstrated level of understanding was coded at Level 3.

Overall, in his model, Chase demonstrated a Level 2 (partial or developing understanding on the standards) scientific understanding. In this sense, he was similar to his classmates at the end of the creek project although many of these models were more complex than Chase's. Students created an average of 5.42 objects (range 1-12), 10.38 factors (range 5-22), and 13.82 relationships (range 5-36). The models addressed an average of 10.6 standards (range 5-16)

Table 5: Summary of expected and observed conceptual understandings along the National Science Education Standards.

Number of conceptual standards in each content area.	Essay		Report		Model 1		Model 2	
	E Expected	O Observed	E Expected	O Observed	E Expected	O Observed	E Expected	O Observed
Standard B - Physical Science	0	0	0	1	0	1	2	2
Standard C - Life Science	2	3	3	9	4	9	10	10
Standard D - Earth and Space Science	2	2	1	3	3	3	4	4
Standard F - Science in Personal and Social Perspectives	4	6	1	6	0	3	5	7
Total number of conceptual Standards	8	11	5	17	7	16	21	23
Mean number of Standards per artifact:		1.29		7.95		2.5		5.15

The summary of expected and observed standards in the standards matrix (Table 5) shows that as a whole, students addressed most of the expected standards and several that were not explicitly part of the assessments. For example, the first model assignment was mapped onto seven conceptual standards but observations of student understandings mapped onto sixteen standards. But the average number of standards that students addressed in an individual artifact was much lower than the expected number in all artifacts except the report (Table 5). The fact that understandings beyond the expectations of the assessments were observed in each artifact emphasis the need for an evaluator to be sensitive to unexpected outcomes when trying to characterize student understandings.

(4) The examination of student understandings across time and artifacts.

The analysis of individual student understandings was accomplished by using the data of standards and levels of understanding described above in step three, 7687 demonstrations of understanding that were entered into a standards matrix. The *Standards* Matrix (shown in part in Table 1) revealed whether or not students' understandings were stable or if (and when) they changed over the semester long creek project. Step 4 analysis was conducted at the individual level of students' demonstrated understandings of each standard throughout the project.

Since some of the measures were completed by individuals and others by groups of two to five students, anassumption had to be made in the analysis of individual understandings. The assumption was that demonstrated understandings in an artifact could be attributed to all authors of that artifact. I.e., if a report provided evidence of a Level 2 understanding about gas solubility, all of the students whose name was on that report were recorded as demonstrating a Level 2 understanding for standard B.2.5. This assumption was based on an ideal situation where co-authors of an artifact negotiate the content representations and through that process develop and refine each other's understandings. The check and balance on this assumption was the pre/post tests and essays, which were completed by individuals. By employing this assumption, it was possible to do a student-by-student, measure-by-measure, standard-by-standard analysis of demonstrated understandings across the semester project using the Wilcoxan Sign-rank test. For example, a student's understanding of the niche concept (Standard C.5.5) could be tracked from pre-test to essay to report to models 1 and 2 and finally to the post-test.

It is important to note that the pairs of variables (n_c) in the sign test consisted of a measure of understanding and the last time understanding on that standard was recorded (L0 values were excluded from the analysis). For example, on standard C.5.5 related to the niche concept, "Jane" may have demonstrated this understanding in all six measures. "Dick" may address it in only four of the six measures (e.g. pre and post tests, report and model 1). Moreover, "Sally's" understanding of this content may only be assessed on the pre and posttest. The sign test for Jane would compare her understanding of competition on the pre-test to the essay, from the essay to the report, from the report to model 1, from model 1 to model 2, and from model 2 to the post test while the sign-test for Sally would only compare pre and post test scores. Thus, the sign test allowed a determination of whether or not scientific understandings were increasing along a particular standard over the course of the semester by looking for signs of positive change within students within particular standards. The outcomes of the sign tests were reported as probabilities.

Early in the semester, most students demonstrated weak conceptual understandings (Levels 1 and 2), a finding that was not unexpected since students had not time to engage in the content. Most of their observations were disconnected, providing evidence that they may have picked up some knowledge of the stream, but had not yet connected these pieces of knowledge into a conceptual framework indicating an understanding of what they observed.

In general, students' scientific understandings along the standards increased over the course of the creek project (p=.05-.000). The pattern of understandings demonstrated in standard C.5.5 (figure 2)is representative of the other standards. As students demonstrated understandings in each succeeding artifact, they showed more connections and higher levels of understandings. Standard C.5.5 shows another common pattern across the artifacts. Frequencies of Level 1 (non-scientific) were low across the artifacts where students chose the content they included (figure 2). This pattern indicates that students may choose to represent content that they understand in their work.

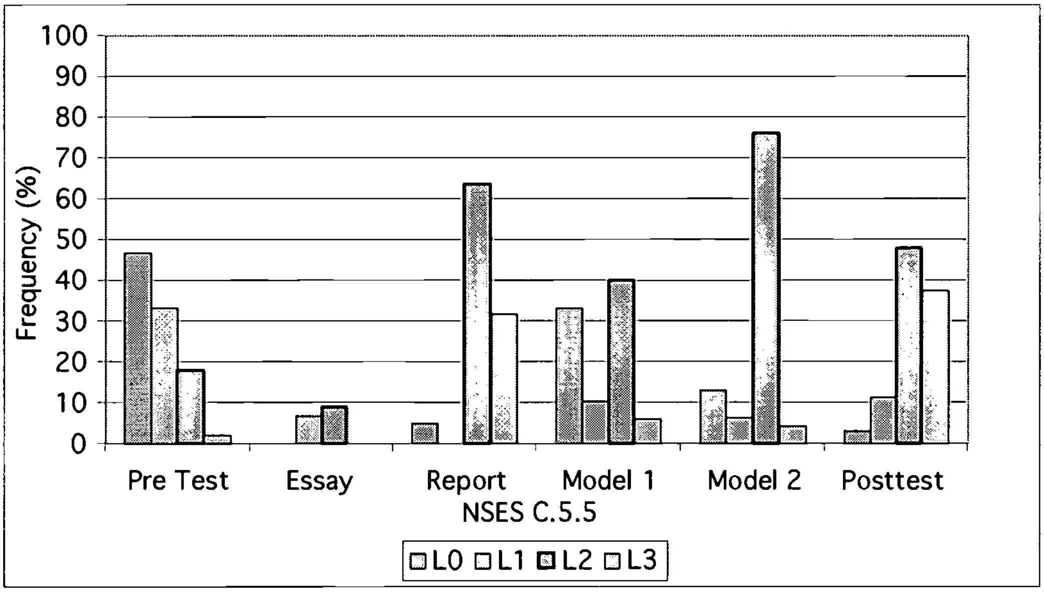


Figure 2: Conceptual Standard C.5.5: the niche concept and environmental tolerances, distribution of proficiencies across assessments. Level 3 (L3) indicates proficient at the *Standards,* Level 2 (L2) partial understanding, Level 1 (L1) non-scientific understanding, Level 0 (L0), no evidence of understandings that map onto the standard.

Several content standards were represented across the set of artifacts although only one, D.3.3 – Interactions in Earth Systems, was expected on all six assessments (Table 3). For the most part, even when a standard was assessed on multiple measures, students often did not display understandings in these same standards. For instance, in the physical sciences (Standard B), only one standard (B.2.5), on the properties of matter - solid, liquids & gases, was expected of and acted upon by students in the pre/post test as well as the reports and two models. However, there were few cases of students attempting this content on consecutive artifacts. For example, of the 23 students that included this content in their reports, only four also included this content in their first model. Moreover, of the 14 students that included this content in their first model, only six included it in their second model. Therefore, the signed-rank test on students attempting this content standard showed no significant differences between levels of understanding demonstrated in the three artifacts.

There was a significant improvement (p = .002) in content representation on standard B.2.5 between the second model and the post-test just a few days latter. Of the 41 students that included this content in the second model, 29 scored higher on the final exam and nine scored lower. Since there was no intervening instruction, a likely explanation for the difference is the different opportunities, expectations, and constraints between these two measures. The intended assessment in the model was very open while the two items (20 & 25) on the test instrument were constrained multiple-choice questions that were scored as correct (Level 3) or incorrect (Level 0). Students also demonstrated significant gains on these two questions ($p=.006$ $n_s=96$) between the pre and post-tests.

Criteria for evaluating the Standards as an Assessment Tool

The usefulness of the *National Science Education Standards* as a tool for describing student achievement across multiple tasks and contexts shows some promise. However, in employing this tool, we encountered several problems that needed to be solved. The problems described in the process part of this paper included: (1) lack of correspondence between content in the curriculum and content in the *Standards*. (2) How to handle different levels of specificity in the assessment criteria (expected content, "if-might" content, and serendipitous content.) (3) Translating content standards into performance standards. Both the promises and the problems need to be considered in order to determine the value of using the *Standards* as an assessment tool.

The standard measurement criteria are validity, reliability, sensitivity and the impact of the assessment on instruction and classroom practices (See Champagne & Newell, 1992; Haney & Madaus, 1989; Kulm & Malcolm, 1991; Malcolm, 1991; Wiggins, 1993). The value and validity of this method have been addressed in a companion paper (Talsma & Krajcik, 2002). Here we consider the criteria of reliability and sensitivity.

Reliability

Reliability is the consistency of the judgment that follows from the use of a measure. Some of the reliability issues in assessment are related to the perceived purpose of assessment. Is assessment simply a tool, like a thermometer, which can be used to obtain some measurement value but with negligible impact on the phenomena being measured, and in which case we would expect high reliability? Or, are there multiple purposes to assessment in which an alternative purpose is to cause students to rethink, to make new links, to ask questions, to build understandings? In traditional reliable assessment practices, similar assessments administered to the same child over time will result in about the same score. But if thinking processes are valued and understanding is conceived as a dynamic process, a learning child should think differently on the second assessment (Champagne & Newell, 1992). According to Perkins (1992), an assessment should be very much a learning as well as a testing experience. Assessments should stretch the learner even as they create an occasion for a learner to display mastery and understanding. Inherently, they test for, and therefore press for, transfer and understanding (Perkins, 1992). Consequently, we would neither expect, nor value, high reliability as traditionally defined when evaluating assessments of scientific understanding.

A second reliability issue has to do with internal reliability. Internal reliability is whether or not performance on different portions of an assessment lead to the same conclusion. The students confronted internal reliability in their creek study as they compared the results of their chemical assessments to the bio-assays. Internal reliability is an issue if we try to apply traditional notions of reliability when considering portfolios of artifacts or student artifacts that may have multiple representations, or themselves multiply represent student understandings. Moss (1994) argues that hermeneutic approaches to assessment can allow students substantial latitude in selecting the products by which they will be represented - a latitude that need not be constrained by concerns about quantitative measures of consistency across tasks.

Wiley and Haertel (1996) offer another means of addressing task reliability without the constraining assumption of homogeneity of tasks. As part of a comprehensive assessment development process, they suggest carefully analyzing assessment tasks to describe the capabilities required for performance, scoring tasks separately for the relevant capabilities, and examining reliability within capability across tasks to which the capability applies. While this approach supports the use of complex and authentic tasks that may naturally vary in terms of the capabilities elicited, it still requires detailed specification of measurement intents, performance records, and scoring criteria. It is this approach that was adapted in the present study, where the assessments (artifacts) employed were analyzed and mapped onto the *National Science Education Standards*.

In this study, reliability was examined by grouping the six different measures in time. For example, the pretest and essay measures occurred at the beginning of the semester, the report and model one at the middle of the semester, and model 2 and the posttest at the end of the semester. With the exception of the pretest and essay assignment, there was no content instruction between the two members of the pair.

A third reliability issue has to do with reliability in scoring, also known as reader or rater reliability. This reliability criteria has a long history in education. Standardize testing evolved and proliferated because the school transcripts became unreliable (Wiggins, 1989). An "A" in a subject meant only that some adult thought the student's work was excellent. However, without being tied to a defined target, there was no possible way to determine what an "A" means in terms of knowledge and understanding.

Alternative assessments, such as evaluating artifacts, include subjective decisions in which rater reliability becomes an important issue. Raters who judge student performance must agree regarding what scores would be assigned to students' work within the limits of what experts call "measurement error." Do raters agree on how an assessment would be scored? Do they assign the same or nearly similar scores to a particular student's work? If the answer is no, then students scores are a measure of who does the scoring rather than the quality of the work (Herman & Winters, 1994). Inter-rater reliability can be improved by careful definition of the relevant information and the use of rubrics similar to those employed in this study.

Sensitvity

Early in the semester (pre-test and essay), the students demonstrated weak scientific understandings, an expected finding since students had not time to engage in the content. Over subsequent assessments, students demonstrated understandings on more standards and at higher proficiencies. However, the sensitivity of the *Standards* tool and four level ordinal coding system affects the number of claims that can be made about changes in understanding.

Sensitivity of an assessment tool is an issue when there is a desire to track changes in understanding over time. The more incremental the changes, the more sensitive the tool needed to be. In order to increase the sensitivity of the standards, this analysis employed a four level ordinal coding scale based on the prior work of educational researchers (e.g. Carey, Evans, Honda, Jay, & Unger, 1989; Grosslight, Unger, Jay, & Smith, 1991; Songer & Linn, 1991; Stratford, 1996). However, in a four level rubric, the intervals between levels are grossly unequal. On any standard, a student might initially demonstrate a Level 2 or partial understanding. On each succeeding measure, they might show more understanding, but never reach the proficient level (3). Likewise, proficiency in the standards, coded Level 3, does not represent highest levels of understandings such as those achieved by experts in a domain.

For instance, Kierra and Magdala built a model about the effects of a culvert on the creek where the length of the culvert affected photosynthesis and respiration rates ("Algae needs sunlight to photosynthesise. A culver would block the sunlight needed from getting to the algae, so the algae would not be able photosynthesise as much"). A steam ecologist may view these concepts through the lens of the River Continuum Concept (Vannote, et al. 1980.) where major bioenergetic influences along the stream are local inputs (allochthonous litter and light) and transport from upstream reaches and tributaries contributing to a mix of hetertrophic and autotrophic sources of energy. A physicist's lens might be on the energetics, with little emphasis on the organisms in which photosynthesis and respiration takes place. A chemist may focus on the reactants and products of the corresponding oxidation and reduction reactions. A cell biologist's understanding might focus on the structure of cellular membranes and the mechanisms by which photosynthesis and respiration occurs in cells. Each of these legitimate differences in perspectives represent sophisticated understandings of photosynthesis and respiration.

Legitimate differences in perspectives and sophistication of understanding will also be evident in individual student's scientific understandings of the natural world, reflecting differences in experience and exposure to science. In a project-based classroom where students pursue different investigations and create different types of artifacts, they may achieve understandings on individual standards far beyond those articulated in the document. A challenge to teachers and others responsible for assessing understanding is to decide how such variability is translated into judgments about the degree to which individual students or groups of them understand the natural world (NRC, 1996, Ch 5). The form of standards assessment used in this study would not capture those higher levels of achievement.

An alternative scale to the four level rubric would be to assess students according to the grade levels differentiated in the standards document. Students' demonstrations of understanding could be characterized as below 4th grade, at 4th grade but not yet 8th grade, between 8th and 12th grade or above 12th grade (= Level 3 or proficient). Such a scale may also seem less abstract to parents and practitioners when discussing student achievement (e.g. Sally shows a 12th grade understanding about the interaction of earth systems (D.3.3) but only a 4th grade level about the properties of matter (B.2.1))

However, a grade-level scale brings into relief a second sensitivity issue, that of specificity of content across the standards. Although the Creek curriculum was relatively balanced between earth science, biology, and chemistry, this balance was not evident when looking at the number of standards addressed as reported in step one of the analysis process. There would probably be little difficulty rationalizing that an 8th grader has achieved a 12th grade understanding of radioactive isotopes (B.1.4). However, claiming that a high school student has only an 8th grade understanding of weather (because standard D.3.2 for grades 5-8 is the highest level at which weather is addressed in the *Standards*) may be problematic when communicating student achievement to communities of teachers, parents and policy makers.

This study also encountered problems in operationalizing the process and nature of science standards (Standards A, E, and G) into forms that would be demonstrable in students' artifacts. For example, in this study, students gave evidence of their epistemological understandings when they reported on the purpose of an investigation or what they intended to demonstrate with their models. Students also exhibited understanding by the way they used evidence in their writing, in supporting an argument or evaluating models based on the use of evidence (table 4). Knowledge about the methods and goals of science traditionally have been treated as declarative knowledge outcomes and measured by objective instruments (e. g., Views on Science –Technology –Society [VOSTS], Aikenhead &Ryan, 1992;Test of Understanding Science [TOUS], Cooley & Klopfer, 1961;Nature of Science Scale [NOSS], Kimball, 1967 –1968;Nature of Scientific Knowledge Scale [NOSKS];Rubba & Anderson, 1978; Science Process Inventory [SPI];Welch &Pella, 1967 –1968). More work is needed to conceptualize how students might represent their epistemologies in artifacts to compliment the other forms of assessment that rely heavily on instrumentation.

Clearly, tool sensitivity, in terms of scalar sensitivity and content specificity, is a criterion for the usefulness of the *Standards* as an assessment tool that needs further development.

Impact of the assessment on instruction and classroom practices

The final criterion of tool usefulness is its impact on instruction. The idea that teachers teach toward the test has become part of the conventional wisdom of education, but has its roots in research (see Kulm & Malcom, 1991; Wiggins, 1989; Wiggins, 1993). The common pattern at the secondary level is for teachers to present the topic, test student to assign grades on the achievement pertaining to the content, and continue on to the next topic (Treagust, Jacobowitz, Gallagher, & Parker, 2001). However, this study did not test the impacts of assessment on instruction. In the study context, the curriculum and artifact production were inextricably bound together by the reforms initiated and implemented as the teachers move toward a more project-based science approach to their science instruction. The impact of standards based artifact assessments on instruction still needs to be established by applying this method to other instructional contexts.

In most such classrooms, students do not get to practice their understandings but instead practice "remembering" (Perkins, 1992). School science tends to present science as a series of known concepts and ideas, a body of knowledge to be mastered (Aikenhead, 1982; Perkins & Simmons, 1988). For example, in an observational study of 11 junior high school science classes, only a very small proportion of tasks required higher-level creative or expressive skills; the predominate activity involved copying information from the board or textbook onto worksheets (Mitman, et al., 1987). Teachers in these classrooms stress correct answers, grades, competition, and public comparison with others. Students are often not provided opportunities to learn the critical thinking skills that permeate the cognitive repertories of accomplished learners (Campione, 1991) and that help develop understandings. This situation is compounded by the nature of instruction in the higher grades, where the emphasis is too often on breadth of coverage. Students are not required to explore a subject in depth, and consequently, it is not easy for them to learn to evaluate new information critically and build the multiple links between concepts that are the hallmark of robust understandings. In the face of such instructional activities, students are likely to conclude that science is static rather than active, and that science proceeds in a linear trial-and-add-new-information approach rather than as a series of conjectures that may or may not be supported (Linn, et al., 1990).

The modes of learning called for in the *Standards* imply markedly different roles and tasks for the students in terms of designing, interpreting, explaining, and hypothesizing. More research is needed of what roles students can play in varied science classroom contexts and they types of work they can produce (Anderson & Helmes, 2001). There is also an ongoing need for research about the intended and unintended effects of assessments on the ways teachers and students spend their time and think about the goals of education (Linn, et al., 1991). It cannot just be assumed that a more "authentic" assessment will result in classroom activities that are more conducive to learning.

The authors recognize one further value of this method of assessment. Because this study characterizes student's understandings, in reference to the *Standards*, as they embark on a three-year program of integrated, project-based science, it provides a foundation for additional research. Interesting questions for follow-up study are, "How persistent are the conceptual understandings developed during the creek study?" "Do student invoke these understandings to make sense of science content in subsequent projects?", and "How do understandings demonstrated in different projects, but mapping onto the same standards (near transfer) compare to the understandings demonstrated in the artifacts examined here?"

The usefulness of the *Standards* as a tool for describing student achievement across multiple tasks and contexts shows some promise in addressing the issues of validity, reliability and impact on instruction. However, when students are learning and developing scientific understanding, the *Standards* are not sensitive enough to capture intermediate changes. Clearly, tool sensitivity, in terms of scalar sensitivity and content specificity, needs further development to meet the usefulness criteria of sensitivity.

By using the *Standards* as a frame of reference, information generated from alternative modes of assessment applied locally can have common meaning and value in the larger community, despite the use of different assessment procedures and instruments in different locales (NRC, 1996 Chap 5). This contrasts with the traditional view of educational measurement that allows for comparisons only when they are based on parallel forms of the same instrument.

References

Aikenhead, G. S. (1982). Science: A Way of Knowing. In V. N. Wanchoo (Ed.), World Views on Science Education (pp. 206-215). New Delhi, India: Oxford & IBH Publishing Co.

Aikenhead, G. S., & Ryan, A. G. (1992). The development of a new instrument:" Views on Science – Technology –Society "(VOSTS). Science Education, 76, 477 –491.

Anderson, C. W., & Roth, K. J. (1989). Teaching for meaningful and self-regulated learning of science. Advances in Research on Teaching: A research annual. J. Brophy, (ed.), 1(1), 265-309.

Anderson, R. D., & Helms, J. V. (2001). The ideal of standards and the reality of schools: Needed research. Journal of Research in Science Teaching, 38(1), 3-16.

Archbald, D. A., & Newmann, F. M. (1988). Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School (1 ed.). Reston: National Association of Secondary School Principals.

Blumenfeld, P. C., Soloway, E., Marx, R. W., Krajcik, J. S., Guzdial, M., & Palincsar, A. (1991). Motivating Project-Based Learning: Sustaining the Doing, Supporting the Learning. Educational Psychologist, 26(3/4), 369-398.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of learning. Educational Researcher, 18(1), 32-42.

Burger, S. E., & Burger, D. L. (1994). Determining the Validity of Performance-Based Assessment. Educational Measurement: Issues and Practice, 13(1), 9-15.

Campione, J. C. (1991). Dynamic Assessment: Potential for change as a metric of individual readiness. In G. Kulm & S. M. Malcolm (Eds.), Science Assessment in the Service of Reform (pp. 301-312). Washington, D.C.: American Association for the Advancement of Science.

Carey, S., Evans, R., Honda, M., Jay, E., & Unger, C. (1989). 'An experiment is when you try it and see if it works': A study of grade 7 students' understanding of the construction of scientific knowledge. International Journal of Science Education, 11(Special Issue), 514-529.

Champagne, A. B., & Newell, S. T. (1992). Directions for Research and Development: Alternative methods of assessing scientific literacy. Journal of Research in Science Teaching, 29(8), 841-860.

Cooley, W. W., & Klopfer, L. E. (1961). Test on understanding science. Princeton, NJ: Educational Testing Service.

Eisenhart, M., Borko, H., Underhill, R., Brown, C., Jones, D., & Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learing to teach mathematics for understanding. Journal for Research in Mathematics Education, 24(1), 8-40.

Goldberg, M. R. (1992). Expressing and Assessing Understanding Through the Arts. Phi Delta Kappan, 73(Apr), 619-623.

Grosslight, L., Unger, C., Jay, E., & Smith, C. L. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. Journal of Research in Science Teaching, 28(9), 799-822.

Haney, W., & Madaus, G. (1989). Searching for Alternatives to Standardize Tests: Whys, Whats and Whithers. Phi Delta Kappan, 70(9), 683-687.

Herman, J. L., & Winters, L. (1994). Portfolio Research: A slim collection. Educational Leadership, 52(2), 48-55.

Huebel-Drake, M., Finkel, L., Stern, E., & Mouradian, M. (1995). Planning a Course for Success: Using an integrated curriculum to prepare students for the twenty-first century. Science Teacher, 62(7), 18-21.

Jackson, S. L., Krajcik, J., & Soloway, E. (1998). The design of guided learner-adaptable scaffolding in interactive learning environments. Submitted to CHI '98.

Jones, R. A. (1985). Research Methods in the Social and Behavioral Sciences. Sunderland, MA: Sinauer Assoc., Inc.

Kimball, M. E. (1967 –1968). Understanding the nature of science: A comparison of scientists and science teachers. Journal of Research in Science Teaching, 5, 110 –120.

Krajcik, J., Blumenfeld, P. C., Marx, R. W., Bass, K. M., Fredricks, J., & Soloway, E. (1998). Middle school students' initial attempts at inquiry in project-based science classrooms. Journal of the Learning Sciences. McCutchan Publishers. 7, 313-350.

Kulm, G., & Malcom, S. (Eds.). (1991). Science Assessment in the Service of Reform. Washington, D.C.: American Association for the Advancement of Science.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, Performance-Based Assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposal to change the nation's testing policy. Harvard Educational Review, 64(1), 76-95.

Malcom, S. (1991). Science Assessment in the service of instruction. In G. Kulm & S. M. Malcolm (Eds.), Science Assessment in the Service of Reform (pp. 187-188). Washington, D.C.: American Association for the Advancement of Science.

Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., & Soloway, E. (1997). Enacting Project-Based Science. The Elementary School Journal, 97(4), 341-358.

Messick, S. (1989). Meaning and Values in Test Validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Mitchell, M. K., & Stapp, W. B. (1994). Field Manual for Water Quality Monitoring (6 ed.). Dexter, MI: Thomson-Shore, Inc.

Mitman, A. L., Mergendoller, J. R., Marchman, V. A., & Packer, M. J. (1987). Instruction addressing the components of scientific literacy and its relation to student outcomes. American Educational Research Journal, 24(4), 611-633.

Moss, P. A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

National Education Standards and Improvement Council (NESIC), (1993). Promises to keep: Creating high standards for American students. Report on the review of educational standards from the Goals 3 and 4 Technical Planning Group to the National Educational Goals Panel (Malcom Report No. Washington, DC: National Goals Panel.

National Research Council (NRC), (1996). National Science Education Standards No. National Academy of Sciences.

Novak, J. D., & Gowin, D. B. (1984). Learning how to learn. New York, NY: Cambridge University Press.

Papert, S. (1991). Situating constructionism. In I. Harel & S. Papert (Eds.), Constructionism: Research reports and essays, 1985-1990 (pp. 1-11). Norwood, NJ: Ablex Pub. Corp.

Perkins, D. N. (1992). Smart Schools: from training memories to educating minds. New York, NY: Free Press.

Perkins, D. N., & Simmons, R. (1988). Patterns of misunderstanding: an integrative model for science, math, and programming. Review of Educational Research, 58(Fall), 303-326.

Perkins, D. N., Crismond, D., Simmons, R., & Unger, C. (1995). Inside Understanding. In D. N. Perkins, J. L. Schwartz, M. M. West, & M. S. Wiske (Eds.), Software Goes to School: Teaching for understanding with new technologies New York, NY: Oxford University Press.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66(2), 211-227.

Rubba, P. A., & Anderson, O. (1978). Development of an instrument to assess secondary school students 'understanding of the nature of scientific knowledge. Science Education, 2, 449 –458.

Schoenfeld, A. H. (1985). Mathematics, technology, and higher order thinking. In R. S. Nickerson & P. P. Zodhiates (Eds.), Technolgy and education: Looking toward 2020 (pp. 67-96). Hillsdale, NJ: Erlbaum.

Schwab, J. J. (1964). Structure of the Disciplines: Meanings and significances. In G. W. Ford & L. Pugno (Eds.), The Structure of Knowledge and the Curriclulum (pp. 1-30). Chicago, IL: Rand McNally & Co

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance Assessments: Political rhetoric and measurement reality. Educational Researcher, 21(4), 22-27.

Songer, N. B., & Linn, M. C. (1991). How Do Students' Views of Science Influence Knowledge Integration? Journal of Research in Science Teaching, 28(9), 761-784.

Spitulnik, M. W. (1995). Students Modeling Concepts and Conceptions: What connections do they make? In Paper presented at the National Association for Research in Science Teaching Annual Meeting. San Francisco. April 21-25, 1995., .

Spitulnik, M. W., Stratford, S., Krajcik, J., & Soloway, E. (1996). Using Technology to Support Students' Artifact Construction in Science. In International Handbook of Science Education Netherlands: Kluwer Publishers.

Stratford, S. J. (1996). Investigating processes and products of secondary science students using dynamic modeling software. Doctor of Philosophy (Education), University of Michigan.

Talsma, V. L. (2002). Student Scientific Understandings in a Ninth Grade Project-Based Science Classroom: A River Runs Through It. Ph.D. Dissertation, University of Michigan - School of Education.

Talsma, V. L. and J. S. Krajcik (2000). Students changing understandings of a stream ecosystem: A Trickle or a Flood. A paper presented at the Annual Meeting for the National Association for Reseachers in Science Teaching (NARST), New Orleans, LA.

Talsma, V. L. and J. S. Krajcik (2002). Assessing scientific understandings: the validity and value of using student artifacts and the *National Science Education Standards* to capture emergent conceptual understanding. A paper presented at the Annual Meeting for the American Educational Research Association (AERA), New Orleans, LA.

Treagust, D. F., Jacobowitz, R., Gallagher, J. L., & Parker, J. (2001). Using assessment as a guide in teaching for understanding: A case study of a middle school science class learning about sound. Science Education, 85, 137-157.

Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R., & Cushing, C. E. (1980). The river continuum concept. Canadian Journal of Fisheries and Aquatic Science, 37, 130-137.

Welch, W., & Pella, M. (1967). The development of an instrument for inventorying knowledge of the processes of science. Journal of Research in Science Teaching, 5, 64-68.

White, R., & Gunstone, R. (1992). Probing Understanding. London • New York • Philadelphia: The Falmer Press.

Wiggins, G. (1989). A True Test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70(9), 703-713.

Wiggins, G. (1991). Standards, not standardization: Evoking quality student work. Educational Leadership, 48(5), 18-25.

Wiggins, G. (1993). Assessment: Authenticity, Context, and Validity. Phi Delta Kappan, 75(3), 200-214.

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring and accuracy. In R. Mitchell & M. B. Kane (Eds.), Implementing performance assessment : promises, problems, and challenges Mahwah, NJ: L. Erlbaum Associates.

[Spitulnik 1998 -]

SE065940

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Comparing Apples & Oranges: Using the NSES as a tool when assessing Scientific understandings	
Author(s): Valerie L. Talsma & Joseph S. Krajcik	
Corporate Source: Paper Presented at NARST Annual Mtg 2002	Publication Date: Apr 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑ Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:	Printed Name/Position/Title: Valerie L. Talsma	
Organization/Address: 6601 Jackson St. Pgh, PA 15206	Telephone: 4124415364	FAX:
	E-Mail Address: Vtalsma@pitt.edu	Date: 26 Aug 02

Vtalsma
4601 Jackson
Pittsburgh PA 15206

ERIC/CSMEE
1929 Kenny Rd
Columbus, OH 43210-1080

Attn! Darcie Hollis

43210+1080

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor: available online http://www.pitt.edu/vytalsma/
Address: papers/NARST2002.pdf
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone:	**301-552-4200**
Toll Free:	**800-799-3742**
FAX:	**301-552-4700**
e-mail:	**ericfac@inet.ed.gov**
WWW:	**http://ericfacility.org**

ERIC/CSMEE
1929 Kenny Rd
Columbus OH 43210-1080
attn Darcie Holles