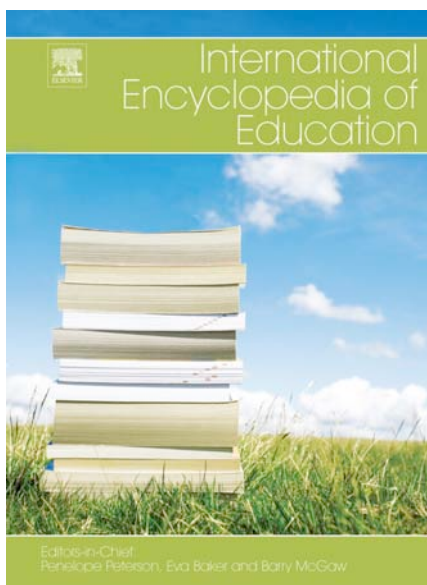


Provided for non-commercial research and educational use.
Not for reproduction, distribution or commercial use.

This article was originally published in the *International Encyclopedia of Education* published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who you know, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

Shemwell J, Fu A, Figueroa M, Davis R and Shavelson R (2010), Assessment in Schools – Secondary Science. In: Penelope Peterson, Eva Baker, Barry McGaw, (Editors), *International Encyclopedia of Education*. volume 3, pp. 300-310. Oxford: Elsevier.

Assessment in Schools – Secondary Science

J Shemwell, A Fu, M Figueroa, R Davis and R Shavelson, Stanford University, Stanford, CA, USA

© 2010 Elsevier Ltd. All rights reserved.

Glossary

Computer-adaptive testing – A method for administering tests that successively selects questions so as to maximize the precision of the test based on what is known about the examinee from previous questions.

Concept map – A network showing the relationships among named ideas (concepts); concepts are connected with labeled arrows that describe the nature and direction of the connection.

Conceptual inventory – A test of student conceptions about the natural world which has been developed from research on student difficulties with particular topics and has been refined, tested, and validated by detailed observations with a robust sample of students.

Conceptual item cluster – A focused set of conceptual questions embedded in a larger assessment. Item clusters are designed to probe student conceptions in detail; they are developed and validated in the same manner as questions for conceptual inventories.

Diagnostic testing – A form of formative assessment usually administered before a sequence of instruction that informs the subsequent planning process.

Formative assessment – An assessment that takes place in the midst of the learning process and is used to guide further instruction and student learning.

Performance assessment – An assessment that involves concrete, goal-oriented tasks in which students manipulate physical objects to try to solve a scientific problem or address a scientific question.

Psychometrics – The field of study concerned with the theory and technique of educational and psychological measurement, including measurement of knowledge, abilities, attitudes, and personality traits.

Question–demonstration assessment – A question and discussion assessment activity involving prediction followed by explanation of a demonstrated science phenomenon.

Reliability – The extent to which an assessment consistently obtains the same scores with the same group of students while they are in a steady state.

Science notebook – An ongoing written account of what students do and learn in their science class.

Scoring rubric – A set of criteria and standards linked to learning objectives that is used to score student performance on a variety of tasks.

Summative assessment – An assessment of learning after it has occurred that is primarily used for reporting the results of instruction to stakeholders in the educational process.

Validity – The extent to which an assessment measures what it is intended to measure.

Introduction

Science assessment in secondary schools ranges considerably in purpose and use, from formative assessment that is indistinguishable from instruction to strictly summative assessment such as standardized testing for accountability. Some assessment tasks lend themselves to both uses, but the differing goals and requirements of formative versus summative assessment result in different design logics and implementation strategies. When formative assessment is melded with instruction, to change assessment is to change teaching directly and vice versa. Yet, both kinds of assessment are inextricably linked to classroom practice. If the school environment, particularly its economy of scale, necessarily shapes science assessment, then assessment – either directly or indirectly – shapes the school environment. Whether formative or summative, the process of assessment – especially the form and content of assessment tasks – signals to students, teachers, and other stakeholders what science in school is supposed to be about.

To assess science learning is to find out what reasoning and actions a student will perform across a range of scientific knowledge domains and situations. In secondary schools, the enduring challenge is for one teacher or proctor to efficiently assess the learning of many students while doing justice to what it means to know and do science. Doing justice to science – the validity of assessment – necessarily involves value judgments and interpretations of what knowing and doing should look like in different situations and across disciplines. While particular content standards vary, most acknowledge that assessed performance depends upon students' knowledge (of theories, facts, concepts, procedures, and strategies), on the one hand, and their acts of reasoning, on the other.

Reasoning and knowledge are inseparable. For example, students who can reason effectively to control experimental variables in one science domain may perform differently in another. Knowledge and reasoning also differ with the social, cultural, historical, and environmental contexts that are reflected or embodied in assessment activities. For example, students' explanations relating force and motion for realistic situations can stand in direct opposition to those they provide when answering textbook questions. For these reasons, valid assessment of science learning requires a wide array of contexts and tasks, a diversity that mirrors the manifold and interconnected ways of knowing and doing that characterize science itself.

In what follows, we briefly describe six categories of science assessment, sampling widely from extant and developing formats and techniques. This sample, while by no means comprehensive, is nevertheless intended to convey some idea of the breadth of available assessment practices in secondary science. In the interest of brevity, we place less emphasis on some time-honored, yet effective, formats such as multiple-choice items and classroom questioning. Instead, we focus on certain advances in assessment that have helped science educators meet the challenge of achieving ever-better validity (doing justice to science) while working within the demands of the secondary school setting.

Question–Demonstration Assessments

Demonstrations of science phenomena, when structured as interactive, question-driven activities, can be powerful tools for probing students' developing knowledge and reasoning. While question–demonstration formats vary, the predict–observe–explain (POE) sequence popularized by Richard White and Richard Gunstone is typical and depicted in [Figure 1](#). In this version, students are presented with the initial conditions of a situation with an uncertain outcome and asked to: (1) predict the outcome; (2) observe what happens; and (3) interpret and explain their observations, reconciling them with their initial predictions.

Question–demonstration sequences are especially useful for formative assessment, and when amplified and extended with discussion and group problem solving, they can form the core of the classroom learning process. They also lend themselves well to larger group settings in which an expert moderator or teacher can intensify engagement, dynamically assess student thinking, and promote and guide the productive exchange of emerging ideas.

Some form of prediction is the signature element of a question–demonstration assessment. While a demonstration's initial conditions will be designed to focus students' attention on specific facts, features, and relationships, the

act of making a prediction brings these ideas into heightened awareness. An important feature of prediction, in contrast to more general forms of questioning, is that students must decide what knowledge applies to the situation at hand. As a result, the act of predicting often elicits students' initial, naive conceptions, making them visible as part of the assessment process.

Observation of the phenomenon provides students with feedback on their predictions. Typically, a teacher or facilitator will perform a single demonstration for the group to observe. Students often observe different things, so teacher and peer mediation are involved in interpreting what happened. With careful moderation, the observation process can stimulate not only productive thinking about science principles but also awareness of the foundations of science in the sense that observed facts are fundamentally interpretations, not mere recordings, of physical reality.

The process by which students explain what they have observed, reconciling these explanations with predictions, is a knowledge building process. When used as formative assessment, a question–demonstration sequence will involve bringing explanations to light while simultaneously assisting students to advance them further. Typical procedures involve short writing prompts extended by discussion and lecture. Responding effectively to students' initial attempts at explanation can demand considerable skill and experience on the part of the teacher, a requirement that stands as a significant challenge to widespread use of this kind of formative assessment. Another limitation is the restricted role allocated to students in the highly scripted question–demonstration process. Although this process may provide for intense engagement and prompt deep reflection, it nevertheless denies students the chance to pose questions and manipulate materials for themselves.

Question–demonstration assessments can be cast in summative form. In one version, students are presented with an initial demonstration setup and asked simply to predict an outcome and justify their prediction; in another, they observe a phenomenon and explain it in terms of science principles and causal mechanisms. These kinds of items are sometimes incorporated into large-scale assessments to signal what is valued in science learning – not only answering questions but also reasoning about and making sense of the physical world.

Performance Assessments

Implementing hands-on tasks in the classroom and in large-scale assessments reflects the importance of doing science. This involves reasoning not just with knowledge in memory but also with an external environment that simulates the conditions and resources with which science is done. Thus, hands-on performance assessment involves concrete, goal-oriented tasks in which students manipulate physical

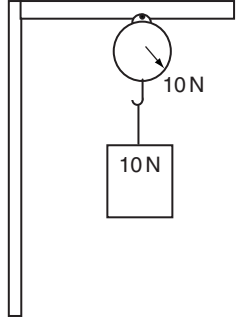
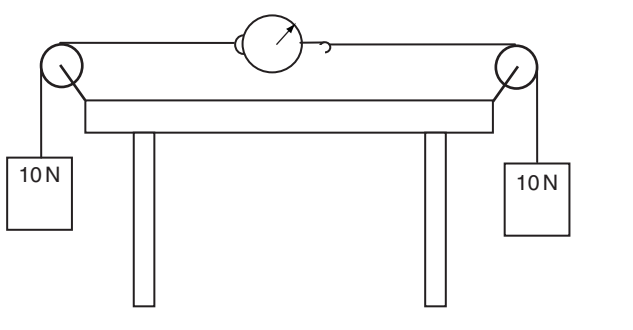
	
<p>Situation I: In this case, the spring balance reads 10 N</p>	<p>Situation II: What will the spring balance read in this case? (ignore the mass of the balance itself)</p>
<p>Part 1: Prediction</p> <p>Which of the choices below best predicts the reading on the balance in situation B?</p> <p>A. 20 N B. 10 N C. 5 N D. 0 N E. 15 N</p> <p>In the space provided, fully explain your reasoning.</p> <p>{A student might predict that the answer is choice A, 20 N, because in the second situation there are two forces, not just one, acting on the balance, providing twice as much force overall.}</p> <p>Part 2: Observation</p> <p>What do you observe? That is, what is the reading on the spring balance in situation II?</p> <p>{The balance in situation II reads 10 N as a result of two opposite 10 N forces. Does the balance in situation I also have 2 opposite 10 N forces? Yes it does!}</p> <p>Part 3: Explanation</p> <p>Why did that happen? Explain the physical principles involved that provide for the reading you observed.</p> <p>{In order for a spring balance to register a force of 10 N and remain at rest, two equal but opposite forces, each of 10 N, must be exerted. Any time there is a tensile force on a static object, an equal and opposite force must also be present.}</p>	

Figure 1 Question–demonstration item in predict–observe–explain format.

objects to try to solve a scientific problem or address a scientific question. The solution or answer is evaluated by a rater or teacher who takes into account not only the student's final result but also the method by which the result was achieved.

Technically, a performance assessment includes a challenge, a response, and a scoring system. The challenge requires students to work with concrete materials to solve a problem and does not specify the steps to be taken. For instance, a student could be given a wire, a bulb, and a battery and asked to light the bulb. The student's response includes his or her actions as well as artifacts produced in the process of tackling the challenge. Responses can be

registered in different formats extending from multiple-choice questions to science-notebook entries. The scoring system delineates the critical knowledge and reasoning expected of students and captures the full range of performance demanded by the task. This could include identification of the right answers, justifiability of procedures, appropriate use of evidence, and effectiveness of problem-solving approaches. [Figure 2](#) shows a physics performance assessment; and [Figure 3](#) shows its task instructions and scoring rubric.

Task demands for performance assessment may be thought of as ranging from knowledge-rich to knowledge-lean and process-open to process-constrained. Some tasks,

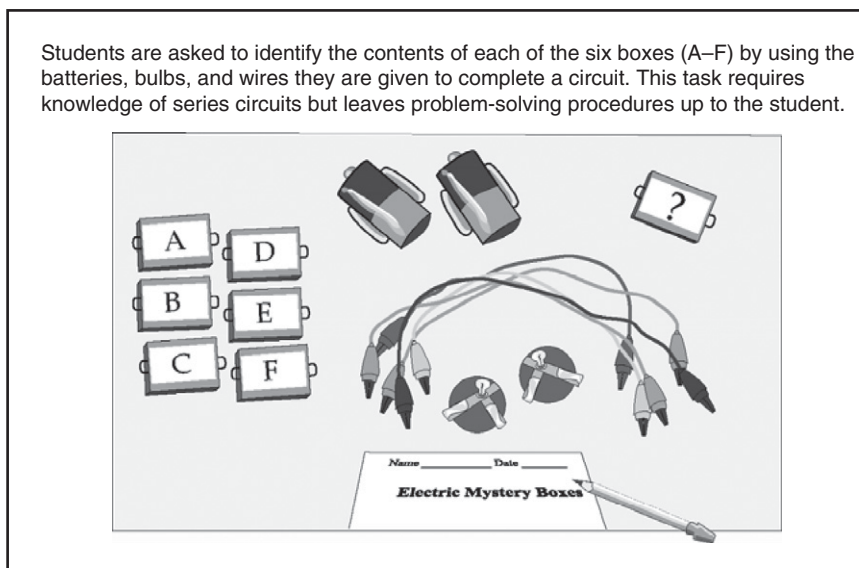


Figure 2 Electric mystery boxes performance assessment. From National Assessment Governing Board. (2006). *Science assessment and item specifications for the 2009 National Assessment of Educational Progress* (pre-publication edition).

The following is a brief description of two warm-up tasks:

1. Students are asked to connect one battery, one bulb, and wires so the bulb lights. They are then asked to draw a picture of this simple circuit.
2. Given mystery box '?,' students are asked to identify whether it contains a battery or a wire. They are told that they can determine the contents of the mystery box by connecting it in a circuit with a bulb.

The following is an excerpt from the main task instructions given to students:

Find out what is in the six mystery boxes A, B, C, D, E, and F. They have five different things inside, shown below. Two of the boxes will have the same thing. All of the others will have something different inside.

[The five options – two batteries, a wire, a bulb, a battery, and a bulb, nothing at all – are presented in words and drawings. Drawings are not provided here.]

For each box, connect it in a circuit to help you figure out what is inside. You can use your bulbs, batteries, and wires in any way you like.

When you find out what is in a box, fill in the spaces on the following pages.

The following is an example of the student response format:

Box A: Has _____ inside.

Draw a picture of the circuit that told you what was inside **Box A:**

The following is a brief description of the scoring system:

For each of the six boxes (A–F), students' responses are scored on two components: (1) identification of the contents of the box and (2) the circuit used to make the conclusion. For each box, if both components are correct, the student receives 1 point; if one or both components are incorrect, the student receives 0 points. Total maximum score is 6 points.

Figure 3 Instructions and scoring guide for electric mysteries task. From Shavelson, R.J., Baxter, G.P. and Pine, J. (1991). Performance assessment in science. R. Stiggins and B. Plake (Guest Eds.), *Applied Measurement in Education* [Special issue] 4(4), 347–362.

like the electric mysteries assessment shown in [Figure 2](#), focus on domain-specific knowledge. Other tasks may focus on more general scientific skills. An assessment of this latter type might ask students to determine which of three paper towels soaks up more water and allow them to design an experiment to answer this question.

Hands-on performance assessments have a wide range of applications from small-scale formative assessment in classrooms to large-scale summative assessment such as the US National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS). Whether in the classroom or on large-scale assessments, the use of hands-on tasks signals the importance of doing science.

Performance assessment in science is challenging. Tasks are difficult to design and costly in terms of materials as well as the time and effort required for administration and scoring. In particular, large-scale assessments require meeting exacting standards for materials and scoring systems. Yet, these challenges are not insurmountable. For instance, the electric mysteries assessment can be used in ordinary classrooms at reasonable cost and with reasonable effort. The use of information technologies to simulate performance tasks can also reduce many of these difficulties.

Information Technologies and Assessment

By transforming the medium in which students carry out science-related tasks, information technologies have the potential to extend the reach of assessment to probe unique aspects of what students know and can do. For example, computers can support interactive models of hard-to-replicate phenomena such as predator–prey interactions. However, information technologies cannot replace other modes of assessment, such as hands-on performance assessment.

As one example of how information technologies can yield rich assessments, the Science Framework for the 2009 NAEP described several types of interactive computer tasks (ICTs), including information search and analysis, empirical investigations, and simulations. While these types are specified in the framework for summative assessment, they illustrate the techniques that potentially apply to both summative and formative purposes.

Information search and analysis tasks echo how scientists and science learners progress by working with the accumulated knowledge of a domain. These tasks provide students with an information database, pose questions, and ask students to find answers by querying the database. Students are assessed on their abilities to select, evaluate, and synthesize information.

Empirical investigation tasks move performance assessments to the computer platform. Doing so can bypass

certain challenges of the hands-on format. For example, computer simulations of experiments can eliminate the hazards of working with certain materials; facilitate manipulation of variables; collect data automatically; and alleviate the costs and logistical complexities associated with procuring, distributing, and storing the physical materials required for hands-on investigations.

Simulation tasks allow students to model, manipulate, and observe scientific phenomena in ways that are difficult with other formats. Some things are not easily seen in real time (e.g., erosion, planetary motion, and chemical reactions) or by the naked eye (e.g., atoms and bacteria), but they can be sped up, slowed down, or magnified in simulations. For example, students could use a simulation of erosion to analyze the effects of various farming practices on erosion rates. [Figure 4](#) shows a computer screenshot from an assessment that asks students to use a model to conduct experiments about population dynamics in a mountain lake ecosystem.

Information technologies can capture a range of student responses. Evidence of students' knowledge, reasoning, and skills can be gathered not only from their final answers but also from the actions that they take while working through an assessment task. On a computer-based task, certain keystrokes and actions can be automatically identified and recorded. Examples of relevant actions include the proportion of time spent on various websites in an information search and analysis, the number of trials performed in an empirical investigation, and the manipulation of parameters for a simulation. Scoring these captured sequences of actions yields an unusually direct assessment of students' strategies for approaching scientific tasks, an important but often elusive aspect of science learning.

Information technologies can also improve the efficiency of assessment. Computer-adaptive testing selects items from an item bank based on a student's responses to prior items. By choosing items that are targeted to a particular student, this technology provides an accurate estimate of individual capability with fewer items. Information technologies can scaffold the administration of complex tasks such as concept maps. Software can machine score students' constructed responses (e.g., essays and concept maps), which is faster and less expensive than human scorers while achieving roughly the same accuracy.

Improved efficiencies can support formative uses of assessment. Information technologies are able to efficiently analyze and summarize the vast amounts of performance data that may easily confound teachers' efforts to understand the state of students' learning. With more information on student performances gathered and summarized more quickly, teachers and students can receive feedback and change courses of action on shorter and potentially more effective timescales. Immediate feedback and interactivity built into information technologies can also directly guide student learning. For example, a

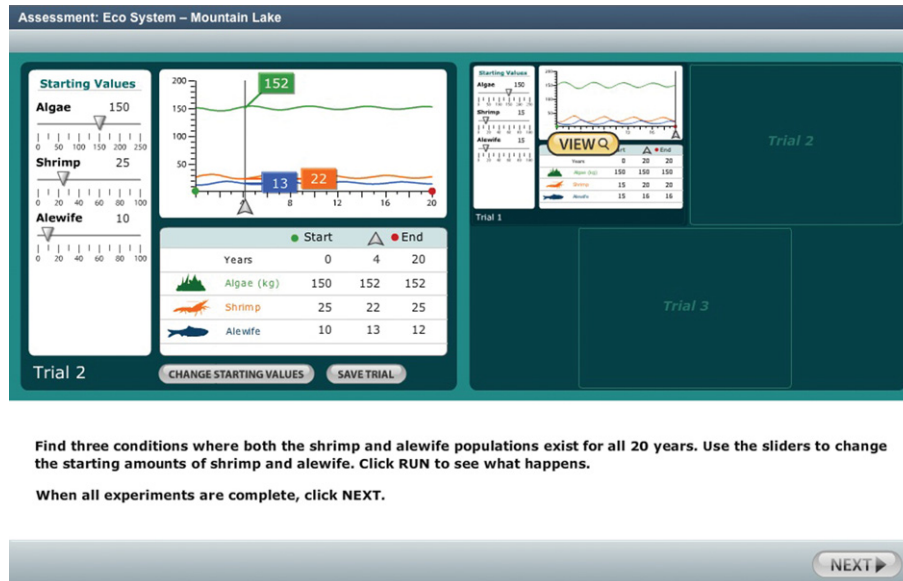


Figure 4 Screenshot from Calipers II predator-prey populations modeling tool. From Quellmalz, E. S., Timms, M. J., and Buckley, B. C. (2009). Using science simulations to support powerful formative assessments of complex science learning. Presented at the annual meeting of the American Educational Research Association, San Diego, CA.

computer-based assessment may immediately alert students if they answer or do something incorrectly, permitting an opportunity to reflect on their understandings and resubmit their answers.

The above examples highlight the promise of information technologies, but significant challenges remain. Unresolved issues include those concerning equity, privacy, and financial and other practical constraints. Further, the use of information technologies in assessment largely represents new psychometric territory. For example, it has not been firmly established whether computer tasks measure the same competencies as hands-on tasks and whether they measure these competencies with comparable reliability and validity. As the use of information technologies in science assessment progresses and expands, these and other challenges will require considerable attention and research.

Conceptual Inventories and Item Clusters

Conceptual inventories and item clusters are developed and used to probe students' knowledge and reasoning about specific science topics in depth. Unlike traditional tests with stand-alone items, these instruments use sets of conceptually related items that allow for deep and explicit investigations of students' particular explanations or mental models of the natural world. The degree of item relatedness ranges in scope from instruments covering broad topic areas to those targeting a single idea or conception. Conceptual inventories and item clusters may be used as tools for diagnostic testing that in turn inform

instructional planning, or researchers and curriculum developers may use them to evaluate students' performances in a randomized trial before and after an experimental intervention. Item clusters probing particular conceptions in depth can be used in large-scale standardized testing. Such clusters provide samples of more detailed information about student learning than typical science achievement items.

Conceptual inventories and item clusters are developed in conjunction with rigorous research into typical naive science conceptions. These instruments generally take the form of relatively short multiple-choice tests, but open-ended questions with detailed scoring systems may also be used. Multiple-choice items such as the one shown in Figure 5 are carefully constructed and validated such that each distractor corresponds to a prevalent conception identified from the research literature. Thus, incorrect responses can reveal as much about students' knowledge and reasoning as correct responses.

Responses to questions such as the one in Figure 5 may seem to suggest that students hold relatively stable and coherent conceptions of the natural world. However, such interpretations should be made with caution, as students' science conceptions are influenced by a range of contextual factors, including the unique sociocultural and affective features of a given situation. The confidence with which an individual's response to an item might be interpreted as reflecting stable knowledge and beliefs therefore depends upon the conditions under which the item was developed and tested.

The Force Concept Inventory is one of the best-known and most widely used conceptual inventories (see Figure 6).

This 30-item multiple-choice test probes conceptions of force and motion using everyday language and semi-realistic situations. Physics students taking this test can often exhibit profoundly nonscientific conceptions despite having mastered more traditional, problem-based assessments. Results like these demonstrate that being able to solve problems does not necessarily entail well-grounded conceptual understanding. This fact, together with continuing research on students' conceptions, has led to a proliferation of conceptual inventories in other science topics. **Figures 7 and 8** show sample items from two of these.

Many conceptual inventories and item clusters are not intended for formative assessment. An exception is a web-based instrument called Diagnoser (see **Figure 9**). This instrument provides students with immediate feedback on their responses to dynamically ordered multiple-choice and short-answer questions. Diagnoser takes more time and asks more questions than most conceptual inventories

What causes day and night?

- A. The earth spins on its axis. (0.66)**
- B. The earth moves around the sun. (0.26)
- C. Clouds block out the sun's light. (0.00)
- D. The earth moves into and out of the sun's shadow. (0.03)
- E. The sun goes around the earth. (0.04)

Figure 5 Example item from the 47-item Project STAR Astronomy Concept Inventory. Some conceptual inventories include proportions of students typically selecting each response option. Here, 26% of students in the research sample responded that day and night are caused by the earth's movement around the sun. From Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching* **35**(3), 265–296.

with items tightly clustered around specific science ideas (e.g., the effect of pushes and pulls, explaining constant speed). The result is an intensive assessment of student conceptions that can guide both teachers and students in their choices regarding further instruction and study.

Conceptual inventories and item clusters define, through the questions they ask, what it means to have a strong grasp of fundamental principles in a domain. These instruments can impact science teachers' views about what should be learned in science and how this should be achieved. Generally, teachers voluntarily select and employ conceptual inventories and item clusters. Students' performances on these instruments confront teachers with the prevalence and sturdiness of students' naive conceptions. By exemplifying essential knowledge and reasoning, conceptual inventories and item clusters illustrate an important way in which science assessment can guide and contribute to teaching practice.

Concept Maps

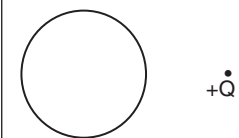
Concepts maps are important tools for measuring the structure of students' conceptual knowledge in a science domain. The integration of ideas, including connections between key concepts, is an important aspect of expert knowledge and a key feature of scientific literacy. As students acquire expertise in a domain through learning, training, and experience, their representations of knowledge begin to more closely resemble the highly integrated knowledge structures that are characteristic of experts.

A concept map consists of nodes and labeled directed lines (see **Figure 10**). The nodes correspond to key terms representing concepts; the lines symbolize a relationship between a pair of concepts (nodes); and the label on the directed line indicates how the concepts are related. Two nodes and a labeled directed line combine to form a proposition, the essential unit of meaning in a concept

- Imagine a head-on collision between a large truck and a small compact car. During the collision:
- A. The truck exerts a greater amount of force on the car than the car exerts on the truck.
 - B. The car exerts a greater amount of force on the truck than the truck exerts on the car.
 - C. Neither exerts a force on the other; the car gets smashed simply because it gets in the way of the truck.
 - D. The truck exerts a force on the car but the car does not exert a force on the truck.
 - E. The truck exerts the same amount of force on the car as the car exerts on the truck.**

Figure 6 Example item from the Force Concept Inventory. From Hestenes, D., Wells, M., and Swackhamer, G. (1992). Force concept inventory. *Physics Teacher* **30**(3), 141–158.

The figure below shows a hollow conducting metal sphere which was given initially an evenly distributed positive (+) charge on its surface. Then a positive charge +Q was brought up near the sphere as shown. What is the direction of the electric field at the center of the sphere after the positive charge +Q is brought up near the sphere?



(a) Left
(b) Right
(c) Up
(d) Down
(e) Zero field

Figure 7 Example item from the Conceptual Survey of Electricity and Magnetism (CSEM). From Maloney, D. P., O’Kuma, T. L., Hieggelke, C. J., and Van Heuvelen, A. (2001). Surveying students’ conceptual knowledge of electricity and magnetism. *American Journal of Physics* 69, S12–S23.

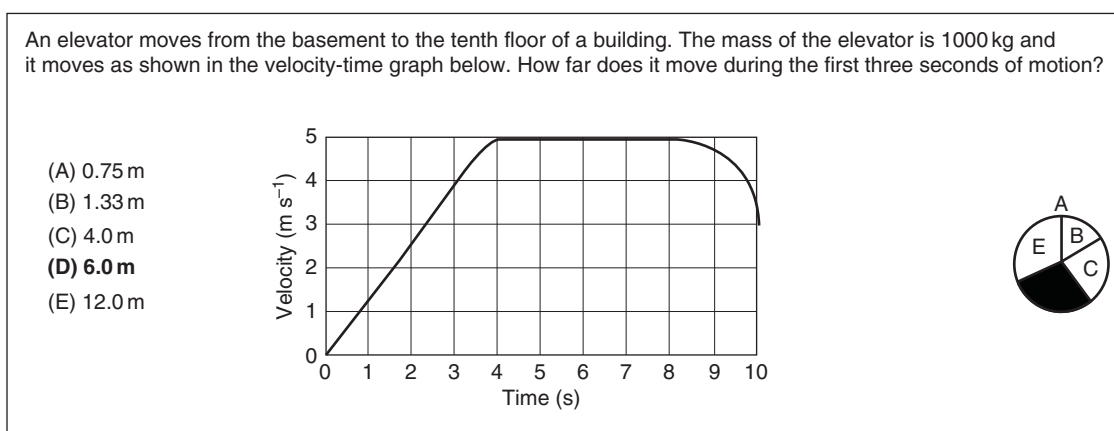


Figure 8 Example item from the Test of Understanding Graphs – Kinematics (TUG-K). The pie chart on the right shows the proportion of students choosing each response option. From Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics* 62(8), 750.

John and his friends watch their radio-controlled car move along a straight path at their school track. John’s friends mark the position of the car as it travels down the track. Some of the data are shown in the table below. Determine the car’s average speed for the time interval shown in the table.

Position (m)	Time (s)
10	2
15	3
18	4
21	5
23	6
24	7
25	8

Type your answer in the box below.
Your answer must be a number.

ms⁻¹

Feedback to response of 3 m s⁻¹:
While the method you used might work in some situations, it will not give you the average speed for the motion unless the object is changing speed uniformly (at the same rate) throughout the motion.

Figure 9 Example item and feedback from Diagnoser. Diagnoser, unlike most conceptual inventories, provides immediate feedback. The correct answer is 2.5 m s⁻¹. From Minstrell, J. (2008). *Diagnoser project: Instructional tools for science and math*. Retrieved February 27, 2008, from <http://www.diagnoser.com>

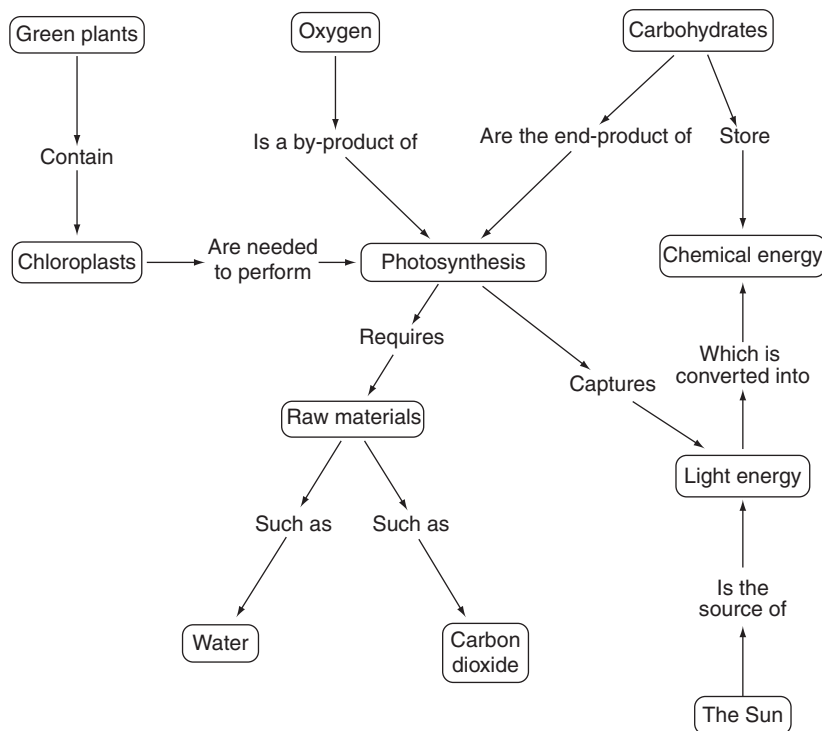


Figure 10 Photosynthesis concept map.

map and the smallest unit that can be used to evaluate the validity of the proposed relationship between two concepts. Concept-mapping tasks encompass a wide variety of techniques that can generally be grouped into two main categories: construct-a-map tasks and fill-in-map tasks. Construct-a-map tasks provide a set of concepts and require students to create all of the nodes and labeled lines in their maps, whereas fill-in-map tasks provide students some or all of the nodes or lines.

In a strict sense, a concept map used as an assessment tool should include a task that draws upon a student's knowledge structure in a domain, a format for the student's response, and a rubric for scoring maps accurately and consistently. Such structured concept-mapping assessments require advanced planning and provide unique challenges in scoring. Scoring concept maps involves evaluating each proposition on the map or comparing students' propositions (two nodes and a labeled directed line) to those on an expert or criterion concept map.

Scoring construct-a-map assessments is generally more challenging than scoring fill-in-map assessments. When students construct their own maps, choices about how to score missing propositions, incorrect propositions, partially correct propositions, and superfluous propositions must all be considered in defining the scoring system. However, while fill-in-map tasks are easier to score than construct-a-map tasks, research suggests that the two techniques do not tap identical aspects of students' understanding:

- Concept terms appear only once on the map
- The map can be organized any way you want
- Use only the concept terms that are provided
- Use only one labeled arrow between two concepts
- You can link a concept to more than one other concept, but you must use separate labeled arrows
- You can only draw arrows between concepts, not to another arrow

Figure 11 Rules for constructing a simple concept map (made available by the Stanford Education Assessment Laboratory); <http://www.stanford.edu/dept/SUSE/SEAL/>

construct-a-map tasks more accurately measure differences in students' knowledge structures.

Training students in the construction of concept maps is an additional challenge of concept map assessments. Concept maps also require students to follow a strict set of instructions (see [Figure 11](#)). Yet, once the process is understood, concept-map assessments can be easily administered to large numbers of students with minimal direction. Many computer programs are now available to assist students in constructing concept maps, and these programs help to minimize the learning curve.

Concept maps can be used effectively as both summative and formative assessments. They appear in science classrooms as homework assignments, small group work activities, full class collaborations, and individual formal and informal assessments. In formative use, teachers sorting through a set of maps can quickly develop an

understanding of prevalent student conceptions. Spontaneous construct-a-map assessments can be assigned in the middle of a lesson, and more structured fill-in-map assessments are often included on major summative assessments in the classroom as well as on large-scale standardized tests. The Science Framework for the 2009 NAEP specified construct-a-map items. It remains to be seen whether this becomes a standard for such items on other large-scale science assessments.

Science Notebooks

Science notebooks are a compilation of entries that provide a partial and time-bounded record of students' classroom experiences. As instructional artifacts generated alongside student activities, notebook entries are tightly linked to everyday classroom learning.

The characteristics of notebooks vary as a reflection of diverse classroom activities and routines. Entries may include defining concepts, identifying relationships, describing experimental procedures, recording observations, and discussing theoretical models. However, science notebooks go beyond writing; they incorporate drawings, data sets, diagrams, graphs, and tables (see Figure 12). These varied forms of representation are essential aspects of scientific inquiry and communication for both students and scientists. Indeed, an important reason for treating notebooks as assessments is to survey students' learning as they engage in a practice that is prevalent among professional scientists.

Scoring criteria for notebooks vary with type of entry. The scoring of an entry on experimental procedures may focus on replicability, while the scoring of recorded observations may focus on level of descriptive detail. When scoring a student's explanation, assessors may focus on the quality of the claim, the type of evidence provided, and the reasoning that links claim and evidence.

Scoring must be carefully aligned to the overall purpose of an assessment. For example, scoring for formative purposes may require looking for particular conceptual difficulties and learning needs. When teachers provide feedback and guidance as written comments, notebooks serve as a valuable record of the ongoing dialog between teacher and student. Teachers and external stakeholders can also score notebooks at the end of an instructional unit for summative purposes.

One challenge of assessing science notebooks is the considerable time and effort required to read and comment on them. This may be ameliorated to some degree by sampling among entries and using tightly focused scoring rubrics. Another difficulty is the requirement to carefully establish procedures and expectations for using notebooks as authentic scientific tools. Failing to do so can result in entries that misrepresent students' true processes of learning or underrepresent their knowledge and skills. These challenges notwithstanding, notebooks provide a unique source of insight into students' knowledge and reasoning in the context of day-to-day classroom activity.

Conclusion

Assessment of science learning in secondary schools is a challenging endeavor. It must encompass a large sphere of activities, many of which can be difficult to render faithfully in a testing environment. It must address complex ways of knowing that are distributed across individuals and specialized tools, and which can shift with changes in context. It must deliver useful information quickly, so that teachers, students, and other stakeholders can readily decide where they stand and what adjustments to make. It must do all of these things at limited cost and with reasonable effort on the part of teachers and proctors.

We have presented here a sample of methods and resources that can begin to address these challenges. In doing so, we have not tried to circumscribe the field but to

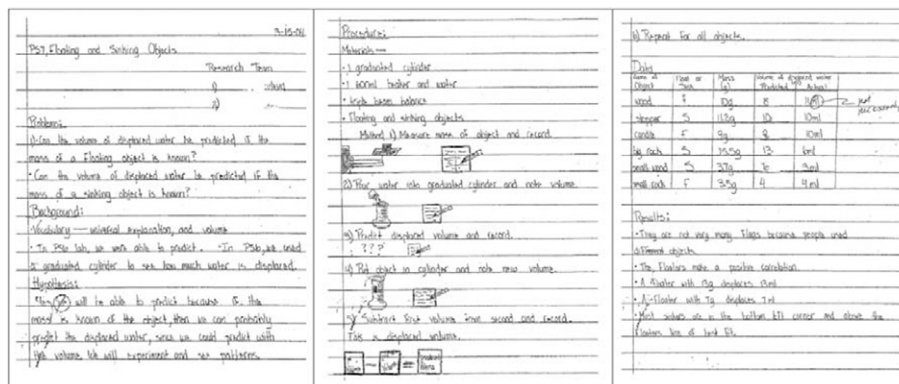


Figure 12 Example of a student's science notebook From Ruiz-Primo, M. A., Li, M., Tsai, S., and Schneider, J. (2007). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. Presented at the annual meeting of the American Educational Research Association, Chicago, IL.

illuminate certain points of progress within it. Such progress is critical if science educators are to close the gap between what is truly valued in school science learning and what comes to be valued because it can be readily assessed.

See also: Assessment in Schools – Primary Science; Formative Assessment; Instructional System Provided Feedback; Portfolio Assessment; Summative Assessment by Teachers.

Further Reading

- Bennett, R. E., Persky, H., Weiss, A. R., and Jenkins, F. (2007). Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project (NCES 2007–466). *U.S. Department of Education*. Washington, DC: National Center for Education Statistics.
- Black, P. (2003). The importance of everyday assessment. In Atkin, J. M. and Coffey, J. E. (eds.) *Everyday Assessment in the Science Classroom*, pp 1–11. Arlington, VA: NSTA Press.
- Halloun, I. and Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics* **53**(11), 1056–1065.
- Halloun, I. and Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics* **53**(11), 1043–1055.
- Hein, G. E. and Price, S. (1994). *Active Assessment for Active Science: A Guide for Elementary School Teachers*. Portsmouth, NH: Heinemann.
- Mintzes, J. J., Wandersee, J. H., and Novak, J. D. (2000). *Assessing Science Understanding: A Human Constructivist View*. San Diego, CA: Academic Press.
- National Assessment Governing Board (2006). *Science Assessment and Item Specifications for the 2009 National Assessment of Educational Progress* (pre-publication edition).
- National Research Council (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Committee on the Foundations of Assessment, Pellegrino, J., Chudowsky, N., and Glaser, R. (eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- Novak, J. D. and Gowin, D. B. (1984). *Learning How to Learn*. Cambridge: Cambridge University Press.
- Pine, J., Aschbacher, P., Roth, E., et al. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching* **43**(5), 467–484.
- Quellmaiz, E. S. and Pellegrino, J. W. (2006). Technology and testing. *Science* **323**, 75–79.
- Resnick, L. B. and Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In Gifford, B. R. and O'Connor, M. C. (eds.) *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*, pp 37–75. Boston, MA: Kluwer.
- Rosenquist, A., Shavelson, R. J., and Ruiz-Primo, M. A. (2000). On the "Exchangeability" of Hands-on and Computer-simulated Science Performance Assessments. *Report*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education and Information Studies, University of California, Los Angeles: US Dept. of Education, Office of Educational Research and Improvement, Educational Resources Information Center.
- Ruiz-Primo, M. A., Li, M., Ayala, C. C., and Shavelson, R. J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education* **26**(12), 1477–1506.
- Shavelson, R. J., Baxter, G. P., and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement* **30**(3), 215–232.
- Shavelson, R. J., Baxter, G. P., and Pine, J. (1991). Performance assessment in science. *Special Issue: Performance Assessment*. Stiggins, R. and Plake, B. (guest eds.) *Applied Measurement in Education* **4**(4), 347–362.
- Solano-Flores, G. and Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice* **16**(3), 16–24.
- White, R. T. and Gunstone, R. F. (1992). *Probing Understanding*. London: Falmer.

Relevant Websites

- <http://cmap.ihmc.us> – CmapTools – Knowledge Modeling Kit.
- <http://ipat.sri.com> – Integrative Technology Performance Assessments.
- <http://www.ncsu.edu> – North Carolina State University, Assessment Instrument Information Page.
- <http://www.sciencenotebooks.org> – Science Notebooks in K12 Classrooms.
- <http://www.stanford.edu/dept/SUSE/SEAL> – Stanford Education Assessment Laboratory.
- <http://www.capsi.caltech.edu> – The Caltech Precollege Science Initiative.