A Strategy for Large-Scale Science Assessment

Penny J. Gilmer
Aaron Rouby
Danielle Sherdan

January 2012

CALA

Center for Advancement of Learning and Assessment

Florida State University, Tallahassee, FL
www.cala.fsu.edu

A Strategy for Large-Scale Science Assessment

Penny J. Gilmer
Department of Chemistry and Biochemistry
95 Chieftan Way
Florida State University
Tallahassee, FL 32306-4390
gilmer@chem.fsu.edu

and

Aaron Rouby and **Danielle Sherdan**
Center for Advancement of Learning and Assessment
210-B Sliger Building
2035 E. Paul Dirac Drive
Florida State University
Tallahassee, FL 32306-2801
arouby@cala.fsu.edu
dsherdan@lsi.fsu.edu

Acknowledgments

This paper was presented at the annual meeting of the Association for Science Teacher Education, in Clearwater Beach, FL, on 6 January 2012.

The work reported in this paper is supported through a grant from Education Research Programs at the Institute of Education Sciences (IES), award number R305A110121, administered by the U.S. Department of Education. Faranak Rohani is the principal investigator for this research. Related information is available at http://cala.fsu.edu/ies/. Findings and opinions do not reflect the positions or policies of IES or the U.S. Department of Education.

Cover photo by "Lab Science Career"

Abstract

Our research seeks to determine the feasibility of a strategy that would expand the range of skills evaluated by large-scale assessment programs as well as enhance classroom instruction via formative assessment. Though the research may have implications for a broader range of subject matter and grade levels, the current research is focused on seventh-grade science content that cannot be assessed adequately with conventional large-scale testing formats.

The research includes three interrelated parts that are designed to make performance measures more cost effective and the results more instructionally useful. The first part involves the feasibility of administering statewide assessments to carefully selected samples of science students to estimate the achievement of groups of students. This sampling strategy will give educators a picture of students' proficiency with skills that cannot be measured through conventional assessments. The second part entails providing teachers with the means of developing parallel assessments to evaluate the achievement of individual students. If the teachers' assessments provide results similar to those obtained from the sampled students, then the two assessments help validate each other. The final part involves training teachers on the most effective approaches for providing constructive feedback to students based on the teachers' assessments (Center for Advancement of Learning and Assessment [CALA], 2011a).

Our research has identified science competencies that go unassessed in typical standardized summative tests in seventh-grade science. Competencies are mental abilities and skills that students must acquire to meet the curriculum standards set within a particular domain at a particular grade level. Competencies refer to students' thought processes that cannot directly be observed, so the degree to which students possess any given competency must be inferred. This inference is based on directly observable evidence that is derived from various tasks designed to assess one or more competencies. A task is a performance assessment, consisting of specific test items, through which we collect evidence about students' competencies.

Introduction

The challenge within science teacher education that this research addresses pertains to large-scale tests used by many U.S. states. As Association for Science Teacher Education (ASTE) members know well, these tests measure only a *nonrepresentative* subset of important science competencies because of the need to test large numbers of students using item formats that can be administered and scored efficiently. Particularly, when evaluations of students, teachers, and schools rely heavily on these tests, teaching and learning tend to emphasize the types of skills measurable by these tests. A consequence is that cognitively complex skills, which also are a vital part of science education, are not emphasized or are absent from typical large-scale tests.

Carefully constructed performance assessments can measure complex science competencies, although extensive use of performance assessments does not lend itself to typical large-scale assessment programs. Since the passage of No Child Left Behind in 2001, most U.S. states use large-scale, high-stakes assessments for K–12 students. These tests tend to use the multiple-choice format almost exclusively, but only some science competencies can be measured effectively using this format. Cognitively complex tasks are less likely to be tested in large-scale tests but could be tested using performance assessments in the classroom (Oosterhof, 2011a, b).

Through our three-year grant from the U.S. Department of Education, we are examining the feasibility of a strategy that may (1) help expand the range of skills evaluated by large-scale assessments in science and (2) add a formative (instructional) aspect to these assessments. The strategy includes three components. The first part involves the feasibility of administering statewide assessments to carefully selected samples of science students to estimate the achievement of groups of students. This estimates student proficiency at the group level, and like the National Assessment of Educational Progress (NAEP, 2011), is not designed to determine individual student proficiency.

The second part entails providing teachers with the means of developing parallel assessments. Using performance assessment specifications, which are essentially recipes for creating performance assessments that adhere to particular criteria, teachers would be responsible for assessing the competencies of their individual students. If the teachers' assessments provide results similar to those obtained from the sampled students, then the two assessments help validate each other.

The final part involves training teachers on the most effective approaches for providing constructive feedback to students based on the teachers' assessments.

Criteria for Science Performance Assessments

Some general considerations for science performance assessments include the following: (1) The performance assessment can be said to be authentic if well-educated people typically recognize the task to be similar to activities and thinking that scientists typically do or that nonscientists are often expected to do in the real world. (2) The assessment presents a task that most students find interesting. (3) Materials and resources required by the performance assessment are commonly available within middle-school science classrooms. (4) Science teachers find the performance assessment to be efficient with respect to resources and time required to administer and score. (5) Clear distinctions can be made between students' appropriate versus inappropriate responses, or between more versus less desirable responses. (6) Although a particular task may result in highly diverse responses from students, science educators will agree on specific characteristics of a correct response. (7) If a particular specification allows for creation of performance assessments within more than one science discipline, the same scoring plan will be applicable to all assessments created from this specification. (8) Prior to seeing the scoring plan, most science teachers would be able to identify specific qualities associated with a student's correct response. (9) Though a given performance assessment may involve combinations of declarative knowledge, procedural knowledge, and problem solving, the scoring of a given performance assessment should focus on only one category of knowledge. (10) Analytical, not holistic, scoring involving a checklist or rating scales should be used.

Theoretical Underpinnings

For our theoretical underpinnings, we use (1) the capabilities-complexity model (Oosterhof et al., 2008) in which we determine the capabilities of the assessment (i.e., declarative knowledge, procedural knowledge, or problem solving) to be addressed first, followed by the complexity level (separated into three levels of complexity for each type of capability), and (2) evidence-centered design (Mislevy et al., 2004) with its *competencies*, *evidence*, and *tasks*. The *competency* model inquires, what collection of knowledge and skills should be assessed? The *evidence* model asks, what behaviors or performances should reveal those constructs? The *task* model questions, what tasks should elicit those behaviors that comprise the evidence?

Identification of Science Competencies That Go Unassessed

Our research has identified science competencies, derived from the Florida Department of Education (2008–2010) Next Generation Sunshine State Standards for seventh grade, that go unassessed in typical, large-scale standardized tests in seventh-grade science. The Florida Comprehensive Assessment Test (FCAT) is the test used statewide to assess these standards (Florida Department of Education, n.d.).

Competencies “represent the set of mental abilities and skills that students must acquire to meet the curriculum standards set within a particular domain at a particular grade level” (Gilmer et al., 2011, p. 2). Since competencies refer to students’ thought processes that cannot directly be observed, the degree to which students possess any given competency must be inferred. This inference is based on directly observable evidence that is derived from various performance tasks designed to assess one or more competencies. A task is a performance assessment, consisting of specific test items, through which we collect evidence about students’ competencies. Therefore, the three concepts critical to the development of an assessment specification are competencies, evidence of each competency, and tasks.

A performance assessment specification provides a recipe for producing a set of parallel performance assessments. Each of our specifications describes evidence that will be used to estimate student proficiency with a particular competency or a series of closely related competencies, and establishes common parameters for tasks associated with performance assessments generated from the specification.

We have developed four assessment specifications this year, and we plan five more for next year, the second year of the grant. Teachers will use these specifications to guide the development of performance assessments to be used formatively and summatively with their students. Similarly, the research team will use the specifications to develop summative performance assessments, which will be administered to students around the same time as teachers administer their summative assessments. Once the scores of the teachers’ summative assessments and the research team’s summative assessments have been tabulated, we will use a statistical technique known as generalizability theory to determine the extent to which the teacher- and researcher-based assessments agree. To the extent that results are similar, this would serve to cross-validate the teacher- and researcher-based assessments.

Our Study

Big Ideas in Seventh-Grade Science in the State of Florida

The Florida Department of Education (2009) organized statewide science concepts into 18 Big Ideas for Grades K–8. Table 1 shows nine Big Ideas in science for seventh grade.

Table 1: Nine of the 18 Big Ideas in Four Body of Knowledge Areas in Seventh Grade in Florida

Body of Knowledge	Big Ideas
Nature of Science	1: The Practice of Science 2: The Characteristics of Scientific Knowledge 3: The Role of Theories, Laws, Hypotheses, and Models
Earth and Space Science	6: Earth Structures
Physical Science	10: Forms of Energy 11: Energy Transfer and Transformations
Biological Science	15: Diversity and Evolution of Living Organisms 16: Heredity and Reproduction 17: Interdependence

For example, Big Idea 1, on the practice of science in the category, nature of science, has four parts:

A. Scientific inquiry is a multifaceted activity. The processes of science include the formulation of scientifically investigable questions, construction of investigations into those questions, the collection of appropriate data, the evaluation of the meaning of those data, and the communication of this evaluation.

B. The processes of science frequently do not correspond to the traditional portrayal of "the scientific method."

C. Scientific argumentation is a necessary part of scientific inquiry and plays an important role in the generation and validation of scientific knowledge.

D. Scientific knowledge is based on observation and inference; it is important to recognize that these are very different things. Not only does science require creativity in its methods and processes, but also in its questions and explanations.

One part of our theoretical underpinning is the evidence-centered design with its competencies, evidence, and tasks. Therefore, in this paper, we have organized the paper around an example of an assessment specification and a performance assessment with its competencies, evidence, and tasks delineated.

Competencies

For this first year of the grant, as a team, we went through the Florida benchmarks for each of these nine Big Ideas for seventh grade and determined whether the *competencies* were testable using the FCAT or not. For example, with the Florida benchmark, SC.7.N.1.1, in the nature of science domain, we developed competencies that are not assessable by FCAT (see Table 2) and then wrote specifications to assess those competencies.

Table 2: Florida Benchmarks Not Assessed in the Florida Comprehensive Assessment Test

Benchmark	**Our Determination of the Seventh-Grade Science Competencies That Go Unassessed by FCAT**
SC.7.N.1.1: Define a problem from the seventh grade curriculum, use appropriate reference materials to support scientific understanding, plan and carry out scientific investigation of various types, such as systematic observations or experiments, identify variables, collect and organize data, interpret data in charts, tables, and graphics, analyze information, make predictions, and defend conclusions.	Student can conduct a scientific investigation consisting of the following subcompetencies: • *Student can formulate a scientifically testable question(s) that relates to the context or data provided.* • *Student can create a plan for carrying out a scientific investigation, including what, when, and how to measure variables.* • *Student can carry out a plan for scientific investigations of various types.* • *Student can organize data by creating a table, chart, or other representation to facilitate interpretation.* • *Student can make inferences and predictions and use the data to defend or refute conclusion.*

Development of Assessment Specifications

From those benchmarks and the list we established of seventh-grade science competencies that go unassessed, we developed four assessment specifications related to these big ideas in seventh-grade science, three in nature of science and one in biological science (CALA, 2011b).

For the benchmark outlined in Table 2, we focused on the competency, "Student can formulate a scientifically testable question(s) that relates to the context or data provided."

Evidence

Evidence of the Competency

We delineate the evidence of the competency as follows (CALA, 2011b):

> Students demonstrate they can develop a scientifically testable question related to variables that are provided to them. Although the variables are familiar to students, the students are not to have had previous experience with generating scientifically testable questions using those variables.
>
> The number of variables provided is limited (i.e., 6–8). All variables pertain to a single scientific context. These variables are *not* to be operationally defined. In other words, variables are expressed in a generic form without specifying how they might be observed, measured, or quantified.
>
> Approximately half of the variables relate to variables in the natural world that can be measured in an **objective** manner, including objects, organisms, events, natural forces, and the like. A competent seventh-grade student should be able to imagine how these variables could be observed or measured without much difficulty.
>
> The rest of the variables, however, pertain to things that typically are observed or measured **subjectively**, such as personal opinions and preferences or pseudoscientific claims (e.g., extrasensory perception, ghosts). These are variables that students typically should avoid when formulating a scientifically testable question. However, a student may use these subjectively observed or measured variables to form a scientifically testable question *if* the student sufficiently operationally defines the variables in order to indicate how they could be studied scientifically.
>
> Objectively and subjectively observed or measured variables are phrased similarly so that students cannot easily distinguish between these variables based on superficial characteristics.
>
> Each student selects two variables from the list. To facilitate scoring, the student is asked to explain how or why each selected variable may be part of a scientifically testable question. For instance, the student might provide an operational definition of the variable in which he or she describes a specific way that the variable could be observed or measured, or the student might describe in an abstract manner why the variable is observable or measurable (e.g., because it is part of the natural world; because humans can sense it via one of the five senses; because it can be observed or measured via some form of technology, such as a microscope, thermometer, etc.).

> The student then writes down a scientifically testable question using the two selected variables. Preferably the student's testable question implies an anticipated relationship between the selected variables. (p. 3)

Scoring Plan in Assessment Specification[1]

Below is the scoring plan that the teachers and external evaluators would use to score the students for evidence of their understanding for assessment specification #1. For all scoring plans, we indicate:

1. "Partial credit is never awarded for any items within the scoring plan, even when multiple points are involved. Credit is to be awarded in full or not at all.

2. Grammar, spelling, or other factors not directly related to the competency being assessed are not reflected in the scoring plan. Using professional judgment, however, the teachers are encouraged to provide students feedback on these additional factors that are not assigned points in this scoring plan" (CALA, 2011b, p. 4).

For this assessment specification #1, to create a scientifically testable question, students choose variables that can be objectively and appropriately quantified. When a student selects from the list a **MORE objective** variable, use the scoring procedure that is **below on the left side**. However, if a student selects more subjective variable(s), full credit is given only *if* the student also describes a specific process that objectively and appropriately quantifies observations. When a student selects a **LESS objective** variable, the scoring procedure that is **below on the right side** is used.

1st Variable

MORE Objective Variable Selected

(Student receives one point for selecting an objective variable)	1
It is obvious from the student's explanation that the student recognizes this variable can be directly observed.	2

LESS Objective Variable Selected

Student describes a specific process to quantify observations of the variable that clearly is both objective and appropriate.	3

[1] This section of the paper is retrieved from http://www.cala.fsu.edu/ies/performance_assessment_specifications

2nd Variable

MORE Objective Variable Selected

(Student receives one point for selecting an objective variable)	1
It is obvious from the student's explanation that the student recognizes this variable can be directly observed.	2

LESS Objective Variable Selected

Student describes a specific process to quantify observations of the variable that clearly is both objective and appropriate.	3

Question Characteristics

Statement phrased as question.	1
Question references the two variables that the student previously listed, AND NO OTHERS.	1
Question implies a relationship (causal or correlational) between the two listed variables.	1

Total number of points for this specification is nine, three for the first variable, three for the second variable, and three for the question characteristics. The scores are added together for each student for each assessment answered by the student.

Task: Performance Assessment Using This Assessment Specification

Once the assessment specification was finalized, the CALA team developed a performance assessment based on the specification. Our sample assessment for this specification is titled, "Formulating Scientifically Testable Questions Related to Flowers" (CALA, 2011b).

The task students are asked to perform for this performance assessment is to develop a scientifically testable question related to flowers, using two of the following variables:

- How beautiful a flower looks
- How often insects (pollinators) visit a flower
- How many flowers there are in a certain area
- How much nectar a flower produces
- How much a person enjoys the smell of a flower
- How much time insects spend visiting a flower
- How much familiarity a person has with a flower

First, students would select and record two variables from the list that they think might have a relationship with each other. Students would hopefully choose variables that they (or another scientist) could actually observe or measure scientifically. Next, students explain why the two variables they chose could be studied scientifically. To do this, students would explain how each variable could be observed or measured (i.e., provide an operational definition). Finally, students would write a scientifically testable question that relates their two chosen variables to each other.

To create a scientifically testable question, students must use variables that can be objectively quantified. When a student selects from the following MORE objective variables listed below, use the scoring procedure indicated earlier in this paper (on the left side).

- How often insects (pollinators) visit a flower
- How many flowers there are in a certain area
- How much nectar a flower produces
- How much time insects spend visiting a flower

However, if a student selects less objective variable(s), the student would receive full credit only if the student also describes a specific process that objectively and appropriately quantifies observations. When a student selects from the following LESS objective variables listed below, the assessor would use the scoring procedure indicated earlier in the paper (on the right side).

- How beautiful a flower looks
- How much a person enjoys the smell of a flower
- How much familiarity a person has with a flower

During the assessment, the guidelines relevant to the administration of summative performance assessments indicate that the students are to be provided no assistance, except for clarification related to instructions, or variables from which students are to develop a research question. Clarification on the variables must not provide information that influences the variables a student selects or the student's eventual phrasing of the research question.

In the first summer of the grant, we shared our four, preliminary assessment specifications with the four seventh-grade teachers working with us. The teachers and an external advisory team provided feedback that guided revisions to the assessment specifications. In addition, we trained the teachers how to use the assessment specifications so that they could write their own performance assessments that they will use in the second semester of this academic year, 2011–12. The teachers were also trained in best practices in formative assessment and in providing formative feedback (e.g., Shute, 2008). The teachers will use their own performance assessments both formatively during the semester and summatively at the end of the semester.

Both the CALA team and the teachers will prepare performance assessments using the same assessment specification to use summatively. The graduate students on the project will administer the performance assessments to the students in these four teachers' classrooms and will score the assessments separately from the teachers' summative performance assessment. We will compare the scores of the summative performance assessments developed by the teachers and the research team to cross-validate each other.

Significance

Dr. Faranak Rohani, the principal investigator, outlined the significance of this research saying, "Given the importance associated with statewide and other large-scale assessments, it is critical that these assessments involve essential skills that are beyond the reach of traditional tests. Furthermore, the central focus of these assessments should be on improving the learning of the students who are taking the tests. Assessments should not be a stratagem employed against teachers, but rather a tool designed to help teachers leverage learning in the classroom" (CALA, 2011a).

Literature

Center for Advancement of Learning and Assessment. (CALA, 2011a). *IES awards three-year grant to CALA*. Retrieved from http://www.cala.fsu.edu/articles/7

Center for Advancement of Learning and Assessment. (CALA, 2011b). *Performance assessment specifications—Phase 1*. Retrieved from http://www.cala.fsu.edu/ies/performance_assessment_specifications

Florida Department of Education. (n.d.). *Bureau of K–12 Assessment, FCAT Science*. Retrieved from http://fcat.fldoe.org/scinfopg.asp

Florida Department of Education. (2008–2010). *Next generation sunshine state standards*. Retrieved from http://www.floridastandards.org/Standards/FLStandardSearch.aspx

Florida Department of Education. (2009). *Science big ideas*. Retrieved from http://www.schoonoodle.com/content_areas/834/subjects

Gilmer, P. J., Sherdan, D. M., Oosterhof, A., Rohani, F., & Rouby, A. (2011). *Science competencies that go unassessed*. Retrieved from http://www.cala.fsu.edu/files/unassessed_competencies.pdf

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/r632.pdf

National Assessment of Educational Progress. (NAEP, 2011). *National Center for Education Statistics*. Retrieved from http://nces.ed.gov/nationsreportcard/

Oosterhof, A. (2011a). *Upgrading high-stakes assessments*. Retrieved from http://www.cala.fsu.edu/files/high-stakes_assessment.pdf

Oosterhof, A. (2011b). Upgrading high-stakes assessments. *Better: Evidence-based Education, 3*(3), 20–21.

Oosterhof, A., Rohani, F., Sanfilippo, C., Stillwell, P., & Hawkins, K. (2008, June). *The capabilities-complexity model*. Paper presented at the 2008 National Conference on Student Assessment, Orlando, FL.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.