Measuring how a college education lays the foundation for "thinking like an expert"

Authors: Louise Yarnall, Larry Gallagher, & Geneva Haertel

December 2009

A college education has long been viewed as a way to improve students' capacity to think critically and argue rationally. Yet cognitive psychologists and philosophers studying the development of general reasoning have found such skills to be profoundly shaped by the depth of students' core content knowledge of the "big ideas" in different domains (Chi, Glaser, & Farr, 1988; McPeck, 1981). Knowing big ideas differs from one's capacity to recall the random facts, procedures, and concepts of any given domain. Rather, studies of expert thinking suggest that "big ideas" serve as schematic organizers of facts, procedures, and concepts, enhancing the efficiency of higher forms of reasoning, such as argumentation, problem solving, and creativity.

Domain-Specific Assessment

SRI International is designing a prototype assessment of how well college students learn these big ideas and reason with them like an expert. This work unfolds against the policy backdrop of increased calls for accountability and rising concerns about the global competitiveness of the American workforce. We call the assessments "domain specific" because they measure schematic knowledge that builds around the non-intuitive, hard-won concepts developed in intellectual disciplines through concerted human effort over time, sometimes many generations. Knowledge of these "big ideas" does not emerge easily from common sense, critical thinking, or even logical reasoning. In this work, SRI draws on the methods and frameworks of cognitive science that examine how domain experts use core ideas and how novices learn them. Educational practitioners rarely employ such methods and frameworks. To build a prototype assessment, the Domain-Specific Assessment project focuses on two popular undergraduate courses: biology and economics.

Many Study, Few Think like Experts

While many American students enroll in these courses, few go on to major or work in the fields, and few can even nominally apply what they learned in school in their everyday lives. The project's working hypothesis is that the problem lies partly in how college students learn and, particularly, in how their learning is tested (Dwyer, Millet, & Payne, 2006; Spellings, 2006). In our view, the machinery of instruction and testing does not consistently focus on helping students learn the "big ideas" of any domain. To be competitive, more American college students need not only to learn these big ideas briefly to pass a test in college; they need to learn how to apply them to real world problems.

This project seeks to develop a new type of assessment that will help post-secondary educators track how well American colleges are teaching students these big ideas and preparing them to apply these ideas in their lives. In this work, the team seeks to document its work for reuse by other educators.

Evidence-Centered Assessment Design

The team uses an evidence-centered design (ECD) approach to assessment, a process that involves systematic documentation of the forms of knowledge to be measured. The team is creating "design patterns" for assessments in undergraduate biology and economics. "Design pattern" is a term borrowed from architecture that refers to the basic design elements that architects use repeatedly in many different building designs. In an analogous way, this project sets out to define the reusable essential elements that test designers can use when measuring student learning in undergraduate biology and economics. It also documents the hypothesized cognitive structures of that knowledge, such as links and connections among different ideas and concepts in the domain.

During the first year of the project from March 2008 to June 2009, the team focused on design and development. This process involved five phases: domain analysis, domain modeling, creation of the conceptual assessment framework, creation of pilot items, and pilot-testing of the items. In the domain analysis and domain modeling phases of the work, the team consulted with experts, including both 4-year college professors and professionals in industry, and community college instructors to elicit core knowledge in each domain. The team carefully documented all decisions about the forms of domain-specific knowledge considered desired undergraduate learning outcomes. The project documentation includes concept maps of core knowledge that are organized according to four knowledge types: declarative *(What is the factual and terminology knowledge for the field?),* procedural *(How do you use specific tools and algorithms relevant to the field?),* schematic *(Why do things happen according to the explanatory ideas of the field?),* and strategic *(When do you apply the different ideas of the field to solve real problems?).*

The team has begun creating the conceptual assessment framework, which involves documenting structures and measurement techniques that may be used to elicit evidence of student performance. In addition, the team generated 16 pilot task scenarios (8 biology, 8 economics) and 128 items (64 biology, 64 economics), and conducted pilot tests and cognitive think-alouds with 77 students and six faculty members to gather data on how the items are functioning. In the coming year, these pilot tests will provide insight into how students progress in learning key concepts in economics and biology. They also will provide ideas about the different ways community college students learn and apply knowledge in the two distinct fields.

The remainder of this report will present the conceptual framework, preliminary findings from the pilot study, and a short discussion.

Conceptual Framework

Setting a higher bar for performance, not remediation, is the purpose of this assessment. The team chose to focus on the first 2 foundational years of community college to assess students who already are in, or heading directly into the workforce, as well as those planning to transfer to 4-year colleges and universities.

American 4-year graduates appear to be falling behind in basic academic skills according to some measures . Measures of domain-specific knowledge are needed because they provide a more accurate portrait of both the depth and competitive advantage that students acquire through American higher education.

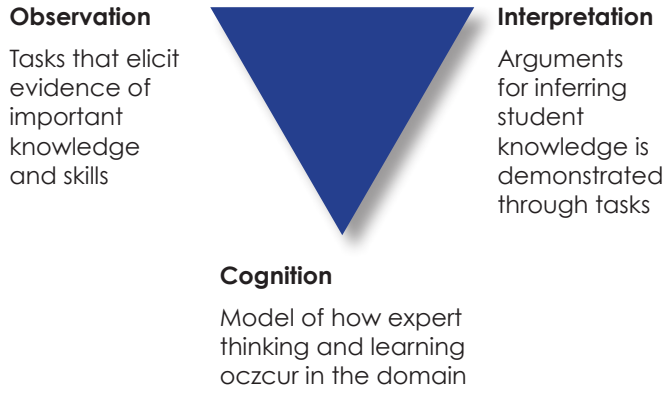
Domain-specific testing requires students to apply the basic content knowledge of a field in complex, expert ways. Cognitive scientists and future employers agree that the hallmark of competent performance is knowing when and why to apply the knowledge, procedures, and strategies unique to a discipline.

There are various models for domain-specific knowledge (Clark & Elen, 2006; Koedinger & Anderson, 1990; Shavelson, Ruiz-Primo, Li, & Ayala, 2003). Yet, there is little, if any, theory or empirical data concerning the way this form of knowledge develops in young adults. Some work has been occurring in the field of biology through the "learning progressions" research, which the team has incorporated into its work (Corcoran, Mosher, & Rogat, 2009). There is even less research in the area of economics around how students develop key understandings of domain-specific knowledge.

This year, the team has conceptualized the domain-specific knowledge outcomes it seeks to measure through the application of ECD and cognitive science theory. We have stayed close to the ECD assessment design framework to design and document all assessment design decisions. In any evidence-based assessment design, three elements—cognition, observation, and interpretation—should be connected and coordinated into a coherent assessment argument that specifies the knowledge to be measured and the forms of evidence that demonstrate such knowledge.

ECD lays out the three core concepts of building assessment arguments from evidence—cognition, observation, and interpretation (see exhibit 1).

Exhibit 1. The Assessment Triangle

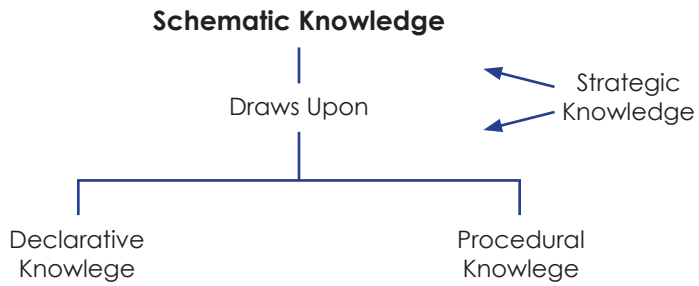


Source: Pellegrino, Chudowsky, & Glaser, 2001, p.44

ECD also identifies five layers of assessment design and implementation (Mislevy & Riconscente, 2006): domain analysis, domain modeling, the conceptual assessment framework, assessment implementation, and assessment delivery. We have completed the first two layers and are currently engaged in the third, as will be described in more detail below. We have based our original formulation of the cognitive structure of domain-specific knowledge based on the work of past researchers in assessment development (Shavelson, Ruiz-Primo, Li, & Ayala, 2003). They set forth four different types of knowledge: strategic, schematic, procedural, and declarative (see exhibit 2).

- Strategic knowledge, as depicted by the arrows below, reflects the capacity to know when and how to apply schematic, procedural, and declarative knowledge. It involves domain-specific conditional knowledge and strategies, such as problem-solving, planning, and monitoring progress.
- Schematic knowledge describes the principles and models that provide explanations for certain phenomena and which organizes declarative and procedural knowledge.
- Procedural knowledge describes the sequential, rule-based activities that lead to expected goals.
- Declarative knowledge describes facts, definitions, and statements of key conceptual relations.

Exhibit 2.
The structure of strategic knowledge



Source: Shavelson, Ruiz-Primo, Li & Ayala, 2003, p.8

Findings

To design the assessments, our team has followed the ECD process. This process has required researchers to iteratively conceptualize how domain-specific content knowledge operates within this cognitive structure. In the paragraphs below, we briefly describe each step in the assessment design process and the resulting insights derived about domain-specific assessment.

Domain Analysis

- We first conducted domain analysis in April 2008. Domain analysis involves gathering information about the domain of interest that has direct implications for assessment: how knowledge is constructed, acquired, used, and communicated. It involves interviewing experts to identify key domain concepts, terminology, tools, knowledge representations, analyses, and situations of use.
- We conducted two 2-day meetings with domain industry and academic experts (5 for biology, 4 for economics) to identify a range of "big ideas," important thinking processes, and knowledge representations central to reasoning like an expert in the domain. This process went smoothly, with experts from industry and academia converging on core ideas relatively quickly. The process was informed by a literature review of research into cognitive science and undergraduate curricula used in each domain.
- The expert discussion led to the identification of the following two levels of biology "big ideas":
  - Fundamental Concepts – Evolution, Bioenergetics, and Systems Biology/Form and Function

- Reasoning Processes – hypothesis generation and testing.

- The economics experts identified three levels of "big ideas":
  - Fundamental concepts – Incentives, Tradeoffs, Efficiency, and Scarcity
  - Models and Relationships – Supply and demand, Inflation and unemployment
  - Modeling Assumptions – Information Availability/Time Constraints and Model Identification and endogeneity

- We developed graphic "concept maps" representing of the core knowledge in each of the two domains. Each concept map interrelated the specific forms of content knowledge (i.e., the "big ideas" listed above) and the different knowledge types (declarative, procedural, schematic, and strategic).

Domain Modeling

- We then conducted domain modeling in the summer of 2008. Domain modeling expresses the assessment argument in narrative form based on information from the domain analysis. It involves specifying the attributes of the assessments: Knowledge, skills, and abilities; characteristic and variable task features; potential work products; and potential observations. In the Principled Assessment Designs for Inquiry (PADI) system, we record all the attributes of the assessment in a document called a "design pattern." A design pattern refers to a set of core assessment characteristics that are robust and re-usable in a domain.

- We conducted three 3-hour working sessions with community college instructors, and invited supportive input from the original domain experts as needed. Instructors discussed the kinds of learning goals they had for their students. They also debated the extent to which the "big ideas" identified by the domain experts would apply to community college courses and to the instructors' specific expertise and course syllabi. One point of tension was that the biology instructors taught microbiology and cell biology courses that typically do not involve direct instruction on evolution, so the instructors expressed doubt about how helpful they could be in providing assessment design input for this topic. There was considerable debate over the next few months as to whether to focus on "evolution" or "the cell as the central unit of life" as "big ideas." We ultimately opted for both "evolution" and "bioenergetics" with a focus on cellular processes. We encountered less debate in economics, where experts and instructors seemed to be in more agreement: Supply and Demand concepts are central to beginning economics courses, and there is some exposure to concepts of Opportunity Cost.

- We documented the decisions using design patterns within the assessment design system. We ultimately refined the concept maps and created a set of three design patterns per domain. Three 3-hour working sessions were held with two biology faculty members and three 3-hour working sessions with three economics faculty members from Foothill College in Los Altos, California. For biology, the design patterns were as follows:
  - Using Biological Scientific Principles to Predict Outcomes

- Using Biological Scientific Principles to Analyze and Explain Current Health and Environment News
- Using Scientific Inquiry Methodology to Analyze and Critique the Design of a Biological Study

- For economics, the design patterns were as follows:
  - Use of Economic Reasoning in Decision Making Situations
  - Reasoning about Market Interactions and Equilibriums Using the Supply and Demand Model
  - Evaluating Government Policies

Conceptual Assessment Framework

- In September 2008, we began developing the conceptual assessment framework (CAF). We are continuing this process currently. In the CAF, we express an assessment argument in structures and specifications for tasks and tests, evaluation procedures, and measurement models. The process involves specifying the operational details about that student, evidence, and task models. We list a range of potential ways to observe and elicit evidence of specific forms of knowledge and skill. We specify rubrics, measurement models, test assembly specifications, PADI templates and task specifications.

Item Creation

- As a first step in item development, we developed prompts to measure declarative, procedural, schematic, and strategic knowledge for specific types of content in both biology and economics. We conducted weekly work sessions from fall through March to develop these assessment items for the purposes of pilot testing. We ultimately developed about 64 items both for biology and economics. We generated eight different scenarios for each domain. We intended each scenario to reflect distinct segments of each domain-specific concept map. Within each scenario, we generated eight items, or four pairs of items dedicated to each of the four knowledge types (declarative, procedural, schematic, strategic). This corpus of items is roughly double the amount we intend to validate.

- We began initial pre-pilot testing in January, having 10 freshmen and sophomores, and four instructors from biology and economics participate in cognitive think-aloud interviews with three biology scenarios and three economics scenarios. Analysis of the transcripts from the initial pre-pilot think-alouds confirmed that our general design approach was sensitive to differences in grade and experience levels. We used instructors and students at two different levels to confirm that there were differences between experts and novices, and also to verify that the high-level performance on these items was consistent with instructors' knowledge.

- We completed the design of the assessment and during March and early April we finalized creation of 16 pilot task scenarios (8 biology, 8 economics) and 128 items (64 biology, 64 economics). This process involved a team of SRI assessment developers with frequent input from domain experts and community college instructors.

Pilot Testing

- We completed pilot test administration in May. For the pilot test, we oversampled and recruited 129 students and ultimately engaged 56 students (57% attrition rate); 40 students took the written versions of the biology and economics tests (20 each domain), and 11 students and two faculty members participated in cognitive think-alouds. For both the written and think-aloud interviews, we sought to sample representative groups of freshmen without any courses in the domain, sophomores who completed one general education course in the domain, and sophomores who completed two or more courses in the domain. We also engaged one instructor from each domain in a cognitive think-aloud interview in going over only the newer scenarios. We paid students with gift cards at a rate of $50 an hour for their participation and paid instructors under a consultancy agreement $50 an hour for their participation. The written tests took roughly 3 hours to administer; the think-alouds took 1.5 hours to administer.

- In June, we began the process of examining the qualities of the student responses to the test items to understand what aspects of our assessment items were working well to elicit evidence of the core learning outcomes we seek to measure, and what aspects were not. We plan to document all of these findings in the PADI system in Task Templates and refine the item prompts, response options, and scoring rubrics based on what we learned from the pilot.

- In the fall, we have refined scoring rubrics by reviewing students' responses in the pilot. This review has led to refinement of partially correct scoring categories. In one type of "partially correct" answer, a student response may indicate deep schematic understanding with an obvious mistake in application. In another, a student response may indicate a specific knowledge deficit or misconception. Most of our discussions focus on how to order different levels of knowledge of the "big ideas." These discussions are leading to refinements to our original design patterns and concept maps that will be available in future reports.

Discussion

From each phase of the design process, we have gained insights into domain-specific knowledge and assessment design.

From domain analysis, we learned that to conceptualize domain-specific knowledge, it is necessary to identify the non-intuitive, hard-won ideas that are central to a field and that routinely inform the ways practitioners in a domain conduct their work. From domain modeling, we learned it is important to define not just the learning outcomes, but also the types of activities in which the knowledge and skills to be learned are most typically applied in a domain. We also learned what kinds of evidence might be provided from different kinds of test items and what features of items and tasks are likely to elicit that evidence. From the CAF process, we have learned the importance of: (1) obtaining specific representations and problems from the field and (2) designing queries for students that seek to measure only the knowledge that is realistic for students to have attained after 2 years of general education in a domain.

Evidence-centered design forces us to ask the critical question, "In each of these responses, what evidence do we have of the core knowledge, skill, or ability being assessed?" As of the summer and fall of 2009, we were learning about how to structure rubrics, using the ideas of Li (2002) to distinguish knowledge types and the ideas of Nagashima et al. (2008) and Wilson (2005) to differentiate different levels in the rubrics.

It has become apparent that the domains of biology and economics have interesting differences in the ways they structure knowledge and assess students. Economics is fundamentally a modeling discipline; from their first days in class, students are learning to extrapolate important elements of everyday situations, and reason using these abstract concepts in well-specified models. Scenario-based assessment tasks in economics, then, appear similar in form to problems that students routinely encounter. Biology, on the other hand, is often taught as a descriptive discipline of different complex systems. "Knowing biology" involves recalling and explaining these systems. Biology students typically have less experience in using their knowledge to reason through a complex "real world" scenario than do their peers in economics classes.

These and other measurement insights are prompting our team to return to the cognitive literature for guidance. When we reach the item validation stage of our work, we are anticipating some interesting, if thorny, questions. For example, does the use of scenario-based assessment introduce a greater proportion of "construct-irrelevant variance" for biology students when compared to economics students? Or, conversely, if we value the ability to apply biological knowledge to the real world, should students be given more opportunity to practice this skill as part of routine instruction than they currently experience?

These and other questions will be explored in future progress reports. We are seeking to present our findings in forms that both researchers and practitioners find useful. We plan to present an overview of the evidence of domain-specific understanding and a listing of the kinds of test item features and tasks that elicit that evidence. In the coming year, the team will be conducting a validation study that includes four key aspects: a correlation substudy comparing our instrument with existing assessment instruments, an instructional sensitivity substudy examining change between freshmen and sophomore cohorts, a cognitive analysis substudy involving think-aloud coding, and an alignment substudy featuring an expert panel's review of how our new assessment tasks—and those on existing instruments—align with the different knowledge types we have defined as central to domain-specific understanding.

Appendix A. Design Pattern Sample

Using Biological Scientific Principles to Analyze and Explain Current Health and Environment News

Title

Using Biological Scientific Principles to Analyze and Explain Real-World Phenomenon

Summary

The design pattern generates assessment tasks that require students to identify relevant biological principles at play in current events related to health and the environment, and then to analyze and critique the reported finding or phenomena using a step-by-step application of the relevant biological principles.

Rationale

Public understanding of current events involving scientific findings and phenomena is enhanced when citizens can link those events to core biological principles. Such a skill permits citizens to think critically about scientific discovery that is unfolding during their lives.

Student Model

Focal knowledge, skills, abilities (grade level implicit)

Ability to identify, among choices, the appropriate biological principle(s) relevant to specific news reports about health and environmental phenomena. (schematic knowledge)

Ability to make explicit the tacit biological principles and reasoning that are relevant to specific news reports about health and environmental phenomena. (schematic knowledge)

Ability to place news reports about biological phenomena into a historical context of scientific discovery about essential biological principles. In other words, students can demonstrate why certain discoveries are of high interest to the scientific community. (schematic knowledge)

Ability to apply the new information in the news report to situations involving one's personal health, matters of public health, or environmental policy. (strategic knowledge)

Knowledge that (declarative knowledge):

Environmental quality:

- The understanding that all life on Earth as we know it adapted because of the capacity of plant cells to, through photosynthesis, convert carbon dioxide into more complex forms and release oxygen from water. The related understanding that when organic matter is consumed or destroyed, the carbon released into the atmosphere in the form of CO2 is generated by fungi, bacteria, and animals consuming or destroying the matter.
- The understanding that the environment is constantly changing and species are adapting to these changes. In natural selection, some members of a population will contain mutations that permit greater survival to these environmental changes. One method of environmental change is by humans, which may occur so rapidly that species do not have time to adapt. The related understanding that evolution of living things is

not teleological, but rather, based on replication/reproduction that lead to an expanding diversity of genotypes, and therefore, phenotypes. Selection pressure acts on phenotypes.

- The understanding that when material goes into the soil, these materials are transformed into other forms in large part because of metabolic processes of organisms.
- The understanding that studying how living things evolve specific functions for reproduction, development, homeostasis, environmental response, and energy consumption can inform the design of new technologies that can improve life.

Personal health:

- The understanding that body functions are based on maintaining cellular health and therefore the health of the organism; exercise and food intake directly influence the life of cells, including their capacity to convert glucose to energy using oxygen. Cells need a continuous supply of energy to perform a variety of constantly occurring functions. The related understanding that all cells in the body are self-replicating and engaged in a continual cycle of life. In other words, the cells we are born with in our body are not those we die with, but rather the descendants of those original cells.
- The understanding that specialized cellular functions are based on both hereditary and developmental factors. These cellular functions can be disturbed any time in the life of the organism because of problems relating to genetics, aging, poor lifestyle, disease or environmental toxins.
- The understanding that healthy cell functioning is dependent upon the ability to maintain homeostatic internal balance and to respond to changing conditions in the surrounding environment. Disturbances in normal cellular function can lead to health problems or death.
- The importance of using drugs judiciously to fight pathogens (e.g., bacteria, viruses) because these are life forms that evolve and may develop a resistance to our drugs.
- The understanding that studying how living things evolve specific functions for reproduction, development, homeostasis, environmental response, and energy consumption can inform the design of new lifestyle, environmental, and nutrition choices that can improve life.

Public health:

- The importance of monitoring how the public uses drugs to fight pathogens (e.g., bacteria, viruses) because these are life forms that evolve and may develop resistance to our drugs.
- The importance of promoting widespread public access to practices and procedures such as hand washing, vaccinations, healthy lifestyle, regular physical and dental checkups to monitor body functions and combat/prevent disease, vector control, maternal health practices, and genetic testing to maintain quality public health.
- The understanding that studying how living things evolve specific functions for reproduction, development, homeostasis, environmental response, and energy consumption can inform the design of new health treatments that can improve life.

Additional KSAs

Familiarity with the underlying declarative knowledge of cellular self-replication processes (i.e., a gene and a protein are not the same)

Familiarity with the underlying declarative knowledge of cellular metabolic pathways for living organisms (photosynthesis) (glycolysis prepares glucose for conversion to usable energy via anaerobic or aerobic chemical processes; anaerobic is typically less efficient than aerobic)

Familiarity with underlying declarative knowledge that genetic mutation occurs in replicating gene sequences and some of those mutations make a species more adaptable to prevailing environmental conditions.

Familiarity with underlying declarative knowledge of osmosis and the basis of exchange through the cell membrane

Familiarity with the declarative knowledge that antibiotics do not work alone but rather, must work in partnership with the host's immune system

Familiarity with underlying declarative knowledge of receptors on cell surfaces to permit delivery of key messages for cellular function and communication. These messages trigger reactions in the cell.

Familiarity with the hierarchical organization of life.

Basic skills of reading and writing.

Ability to interpret and analyze tabular and graphical data?

Basic computational and arithmetic skills.

Understanding the steps of the scientific method.

Evidence Model

Potential observations (student actions)

(How would we recognize the focal KSAs when we see them?)

Quality of student identifying the correct biological principle(s) in play in a news report.

Quality of providing sufficient detail and accurate sequence to explain how the biological principle(s) function in the news report.

Quality of linking a finding in the news report to biological dogma and principles to characterize how the finding advances prior scientific knowledge. (e.g., so what?)

Quality of generating a set of appropriate new personal or public policy practices that apply the findings in the news report, or, by contrast, being able to explain why the news report does not necessarily lead to any changes in existing personal or public policy practice.

Quality of recognizing and correcting a common misconception

Potential work products (artifacts)

Multiple choice question (e.g., linking the news finding to appropriate biological principles)

Short answer response

Written explanation, with diagrams as needed, to illustrate steps in a biological process

Argument advocating for or against a specific public health or environmental policy recommendation, citing relevant theory from biological principles and evidence from news report

In class, I might have the students develop something like a public health pamphlet, however, this wouldn't be appropriate for the time constraints of the assessment at hand

Potential rubrics

3 – Student identifies correct biological principle(s), provides elaborated step-by-step description of underlying biological process, and describes the specific scientific significance (or lack thereof) of the reported finding

2 – Student identifies correct biological principle(s), provides a generally correct sequential description of the underlying biological process, and provides a generally correct characterization of the scientific significance (or lack thereof) of the reported finding.

1 – Student correctly identifies biological principle(s), but may not provide a generally sequential description of the underlying biological process, and does not provide a generally correct characterization of the scientific significance (or lack thereof) of the reported finding.

0 – Student fails to identify the correct biological principle.

Task Model

Characteristic task features

Task must include a news report of either a health or environmental phenomenon OR a scientific finding relevant to health or the environment (additional layer of difficulty: maybe item isn't so obviously tied to health? E.g. discovery microbes, or discovery of DNA. Or maybe this would be too difficult)

Task must require students to apply biological principle(s) and either: (1) characterize the relative significance of the news for biological science, or (2) describe the changes in personal or public policy practice that logically flow from the news or (3) do both.

Variable task features

Familiarity of the topic of the news report.

Number of relevant biological principles

Level of technical detail required to explain the relevance of the news

Genre of information presentation (e.g., a mass media report or a popular scientific journal report or an email from a relative/friend or a claim published on a commercial product)

References

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise.* Hillsdale, NJ: Erlbaum.

Clark, R. E., & Elen, J. (2006). When less is more: Research and theory insights about instruction for complex learning *Handling complexity in learning environments: Theory and research* (pp. 283–295).

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform.* Philadelphia, PA: Consortium for Policy Research in Education.

Dwyer, C. A., Millet, C. M., & Payne, D. G. (2006). *A culture of evidence: Postsecondary assessment and learning outcomes.* Princeton, NJ: Educational Testing Service.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14*(4), 511-550.

Li, M. (2002). *A Framework for Science Achievement and Its Link to Test Items.* Stanford, Palo Alto.

McPeck, J. E. (1981). *Critical thinking and education.* New York: St. Martin's Press.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Nagashima, S., Brown, N. J. S., Fu, A., Timms, M., & Wilson, M. (2008). *A framework for analyzing reasoning in written assessments.* Paper presented at the Annual Meeting of the American Educational Research Association.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment.* Washington, DC: National Academy Press.

Shavelson, R., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2003). *Evaluating new approaches to assessing learning.* (No. CSE Report 604). Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation (CSE).

Spellings, M. (2006). *A test of leadership: Charting the future of U.S. higher education.* Washington, D.C.: U.S. Department of Education.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum Associates.