- - - D R A F T - - -

INTRODUCTION TO THE BIOMASS PROJECT:
An Illustration of Evidence-Centered Assessment Design and Delivery Capability

Russell G. Almond[1], Andrew B. Baird[1], Cara Cahallan[1], Howard Chernick[2], Louis V. Dibello[1], Ann C.H. Kindfield[3], Robert J. Mislevy[4], Deniz Senturk[5], Linda S. Steinberg[6], and Duanli Yan[1]

[1] Educational Testing Service
[2] Fujitsu Consulting
[3] Educational Designs Unlimited
[4] University of Maryland & Educational Testing Service (co-Principal Investigator)
[5] Educational Testing Service intern, Summer 2000
[6] University of Pennsylvania & Educational Testing Service (co-Principal Investigator & Project Manager)

ACKNOWLEDGEMENTS

The Biomass project was funded through Assessment Futures, a joint project of ETS and the College Board, from its inception in January 2000 through June 2000. It was supported by ETS Research from July 2000 through its completion in September 2000. Our subject matter expert consultants were invaluable in working through the issues of standards, claims, and evidence that underlie the project, and in offering suggestions along the way for the prototype. They are Dirk Vanderklein, Scott Kight, Cathryn Rubin, Sue Johnson, and Gordon Mendenhall. For providing data on early field trails of Agouti Segment 1, we thank the ETS Summer 2000 Interns, the Weston scholars at Montclair State University and their advisor Prof. Lynn English, and Russell's buddies at the Knight Dreams comic book shop. Dr. Mislevy received additional support under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

0. INTRODUCTION

This chapter describes the design rationale for a prototype of an innovative assessment product, and the process that led to the design. The goals of the Biomass project were to demonstrate (1) a assessment product designed to serve two new purposes in the transition from high school to college, and (2) the capability needed to produce this kind of assessment product. The conceptual design framework for the project is "evidence-centered assessment design," or ECD for short. This presentation describes the processes by which we designed the Biomass prototype within this framework. We discuss their importance in terms of the design objects and delivery system components, and justify the choices we made as design decisions that serve the product's purposes, in light of the constraints and affordances we have assumed. Fuller discussions of the ECD design process and delivery architecture can be found in Mislevy, Steinberg, and Almond (in press) and Almond, Steinberg, and Mislevy (in press).

Section 1 of the chapter expands on the goals of the project. Section 2 constitutes the bulk of the chapter. It describes the sequence of design activities that led to the requirements for the assessment product, in terms of the final collection of assessment design objects and processes of the assessment delivery system. Examples from the prototype show how these requirements were realized in specific assessment elements. The project led to a working prototype of an assessment product, but one constructed on the lab bench. However, the collections of design objects constituting various stages of product design, as well as the delivery system architecture, and the design process through which the prototype was developed, are meant to be scalable.

1.0 GOALS OF THE PROJECT

The goals of the Biomass project were as follows:

- To create a prototype of an assessment that is fully functional, though with abbreviated content, that demonstrates a kind of assessment product that supports standards-based learning.
- To provide a process for giving functional meaning to authoritative standards, in terms of the expectations they imply for what students should be able to do in what kinds of situations.
- To illustrate how the results of this process lead to the elements of an evidence-centered assessment design framework and a compatible assessment delivery system.
- To highlight aspects of this prototype's development that can be scaled up, in terms of (a) *processes* for moving from initial information about the domain and product requirements to design objects, and (b) *capabilities* for supporting operational task development and assessment delivery.

1.1 Purposes of the Assessment Product

The Biomass project was begun under the ETS/College Board Assessment Futures initiative, "Assessment in the Service of Transitions." This initiative embraced a vision of lifelong learning, in which students flow in and out of the stream of education. Assessment to serve this kind of learning is no longer a matter of a few high-stakes selection or placement tests at fixed and predictable points. The challenge is to guide and certify learning that may occur at many times and places, and can be used in different ways by different people--teachers and parents, college admissions officers and employers, and students themselves.

Assessment in service of transitions can encompass assessments that serve a variety of purposes, from high-stakes certification tests and curriculum-based achievement tests, to diagnostic practice tests and intelligent tutoring systems. We chose to focus on two particular purposes in the high-school-to-college transition, partly for its own sake; it is a familiar context, yet it offers great opportunities for innovation. What's more, it allowed us to demonstrate an approach that can be used for assessments that serve purposes which cut across many transition contexts. Thus, we aimed to demonstrate (1) a prototype assessment product designed to meet new purposes in the transition from high school to college, and (2) the more-broadly applicable capability needed to produce this kind of assessment product.

The demonstration product can be used in two ways, which address complementary purposes that concern the same transition: from a learning context in a standards-based curriculum (e.g., high school class, on-line course), to a position that requires evidence of whether the standards have been met (e.g., college admissions, course placement, job selection).

Culminating Assessment. The product is first designed to support use as a culminating assessment. It measures and reports on students' learning in terms of authoritative, ambitious, standards in a particular domain (grades 9-12), along with supplemental information designed to provide direction for improving learning. This supplemental information is intended for students, teachers, school administrators, and college admissions personnel, to prepare for and to pass on information about the high-school to college transition. Web-based presentation is required.

But it is becoming increasingly clear that one cannot obtain information about students' mastery of ambitious standards with a 'drop-in-from-the-sky' assessment—that is, a test administered to students without knowledge of their instructional background, including the content, methods, representational forms, or types of assessment they have experienced. As we shall see in the case of science standards, getting evidence for standards that deal with high levels of inquiry and model revision can require complex observations in complex situations. Taking statements of standards seriously means constructing tasks with terminology, formats, expectations, and representational forms that are not universally familiar to students.

Interim Assessment. So that students can become familiar with the forms and conventions of the culminating test, as well as the content and expectations, the product is also designed to support use in a self-evaluative mode. In this use it informs students and teachers about progress toward mastery. It addresses the same knowledge base and skills as the culminating assessment, but it would be used by students working individually or together, often as part of a course, to help practice and prepare for the culminating assessment. Again, Web-based presentation is required.

Feedback in both uses is designed to support learning in the domain, and thereby improve understanding and performance. The feedback in the self-evaluation use is finer-grained and more targeted. In both cases, the feedback is intentionally independent of specific instructional approaches or curricula.

1.2 Relevance to Standards-Based Assessment

As mentioned above, the Biomass prototype assessment is intended to be "standards-based." Standards of all flavors are au courant in education today: Content standards, delivery standards, and performance standards to name a few. As of this writing, forty-nine of fifty states (Iowa is the holdout) have learning standards of one kind or another in at least some content areas, for at least some grade levels. Under the recently signed No Child Left Behind legislation, all states will be required to do so to have standards and assess them annually, beginning in Reading and Mathematics and expanding by 2006 to Science. The National Council of Teachers of Mathematics published content standards for teaching mathematics (NCTM, 1989), the American Federation of Teachers describes how to "define world class standards" (AFT, 1995), and the New Standards project defined performance standards for high school English Language Arts, Mathematics, Science, and Applied Learning (New Standards Project, 1997).

Despite all of these standards, however, there is surprisingly little agreement about exactly what defined standards mean, or how one goes about assessing students' performance in their light. Standards documents typically mix descriptions of what student know, what they do, and tasks they might be administered. As a result, much effort is currently being expended on solving problems such as characterizing how well a given set of tasks maps into a set of standards, or what level of performance should be designated on a given test as meeting the standard.

The approach to standards-based assessment that is described in this chapter moves from (possibly multiple) statements of standards in a content area, through statements of the claims about students' capabilities the standards imply, to the kinds of evidence one would need to justify those claims, and finally to the development of assessment activities that elicit such evidence. These steps require working from the perspectives of not only experts in the content area, but experts in teaching and learning in that area. In this way, the central concepts in the field and how students come to know them can be taken into account. Moreover, we incorporate the insights of master teachers into the nature of the understanding they want

their students to achieve, and how they know it when they see it. This is the foundation of knowledge needed to design a coherent system of standards and assessments.

We argue that to achieve replicability and scalability, we must expend this effort up front, working through the connections between claims about students' capabilities and production of evidence in situations that bear certain features. In this way we move beyond thinking about individual tasks, to seeing tasks as instances of prototypical ways of getting the same kinds of evidence about aspects of knowledge. We attain a better conceptual grounding of these aspects of knowledge than simply 'tendency to do well on certain kinds of tasks.' Approaching the problem in this manner increases the prospects of recognizing aspects of knowledge that are similar across content areas or skill levels, and similarly being able to re-use schemas for obtaining evidence about such knowledge as it specializes to different particulars. In this way we make explicit just what evidence we require to justify the claims we want to make about students; we force agreement on the central issue of evidence from the very beginning of the design process.

As we shall see, we are able to take advantage of developments in technology and statistical modeling to extend the range of evidence we can collect and interpret, and thus expand the universe of claims we can support. Technology and statistical models do not drive the design process, however. They open up possibilities and they provide affordances—but they do not, in and of themselves, tell us what it is we want to know about students' knowledge. We use them not simply because they are available, but only as they serve our proposes. That is, the technologies help us elicit evidence of the targeted knowledge and provide useful information to people who use the product or the feedback it provides (Messick, 1994).

1.3 An Approach to Assessment Design

Designing educational assessments presents us with the same kind of challenge as designing buildings, bridges, and airplanes: It is a problem of *design under constraints* (Descotte & Latcombe, 1985, Katz, 1994). Certain foundational principles, such as Newton's laws and properties of materials in engineering problems, must be obeyed. They do not specify a design but they limit the space of viable possibilities. Constraints arise from funding, equipment, deadlines, legal requirements, and available personnel. There are purposes to be served—often multiple purposes, addressing the needs of different users. A successful design must accord absolutely with foundational principles, and it must satisfy constraints and meet purposes as best it can, within the constraints it must satisfy, with the resources it has available.

The foundations of assessment design are twofold: Principles of reasoning from evidence, which cut across time, place, and disciplines, and, for the domain of interest, a conception of knowledge and ways to get evidence about it.

The principles of evidentiary reasoning concern how one reasons from uncertain and particular observations of what students say and do in a handful of specific circumstances, to inferences about what they know or do, or feedback about what they might do next to learn more. These principles are like Newton's broadly applicable laws, as they specialize to, say, designing houses. An assessment that does not address these issues risks failure in several ways, in each case because inferences based ostensibly on assessment data are not supported. In familiar assessment terminology, the resulting inferences will be invalid, unreliable, or unfair.

Conceptions of knowledge and skill in the domain, and what can be observed as evidence of these skills, are like the nature of the building materials one uses to build a bridge, or the geological composition of the land that will support the bridge. These considerations are more particular than Newton's laws, but they are fundamental for the job at hand.

Given these constraints, the designer must determine the individual elements of a specific design. Many design choices must be made, among possibilities that are all consistent with the fundamentals, but may satisfy users' needs better or worse, at higher or lower costs, and employ different materials or hold implications for future use.

Enough experience with bridges and buildings has accumulated for engineers to recognize patterns that satisfy fundamental constraints, to jump-start designs that address particular purposes and local constraints. In architecture, for example, CAD (computer aided design) programs take many physical and material

constraints into account automatically. They provide tools that simplify recurring classes of constraints--modern houses will have plumbing, electrical, and HVAC systems, for example, and they will be used by humans who are five or six feet tall and need places to sleep, eat, and store their possessions.

In assessment, understandings have grown about certain forms and practices of assessment that work well for certain recurring purposes and constraint situations. Familiar configurations such as oral examinations to gauge subject-matter learning date back to medieval universities, for example, while multiple-choice standardized tests first appeared in the Army Alpha intelligence test in World War I. Schemas for the embedded assessments in intelligent tutoring systems have been proposed and studied only more recently.

The "evidence centered design" (ECD) framework for analyzing, designing, and implementing assessments is an attempt to explicate cross-cutting patterns of evidentiary reasoning as they are used in all of the familiar assessments described above, as well as new kinds of assessments that serve unfamiliar purposes or use novel methods of data collection. The framework is defined to capture recurring elements and relationships that satisfy foundational constraints. The elements are defined at a level of generality that accommodate assessments across purposes, content domains, and delivery systems. The goal of designing a particular assessment, to meet particular purposes and satisfy particular constraints, can then be cast as an exercise in specifying the elements of this general model.

This presentation describes the processes by which we designed the Biomass prototype within this framework. We discuss their importance in terms of the design objects and delivery system, and justify the choices we have made as design decisions that serve the product's purposes, in light of the constraints and affordances we have assumed.

2.0 DESIGNING *BIOMASS*

The overall process of designing BIOMASS consisted of the iterative gathering and organizing of information. Different iterations involved changes in sources of information, the manner in which the captured information was represented as part of the design, and the audiences for the design we were producing. As we moved from analysis of the subject matter domain and prototype requirements, through use of this information in sketches of assessment content and processes, to the final complete and integrated specification of the prototype's operational components, the grain-size and nature of the data gathered became increasingly refined and more technical. The shift in the focus of information-gathering required a corresponding shift in the design objects used to express it.

Regardless of design phase, however, the manner in which all aspects of the design process were carried out was guided by the requirements of evidence-centered design. That is, assessment tasks, in the contexts of both the domain and product requirements, are always viewed in terms of the opportunities they represent for eliciting some body of explicitly defined evidence agreed to as required for making specific valued and valid claims about students. Section 2 is organized in terms of the successive iterations of gathering, organizing, and refining information about the domain and the intended product into drafts of assessment design objects. Figure 1 depicts the evolving structures of an educational assessment as it iterates through the design process, from inception to delivery.

As the first steps, it was necessary to choose a subject matter domain and a group of domain experts. Data gathering then proceeded with laying the substantive foundation for the prototype: selection and use of pertinent educational standards, selection of illustrative topics within the subject matter domain, definition of relationships between standards and subject matter topic content. After that the focus shifted, in turn, to the claims we would want to make about students as a result of their performance in this prototype assessment, what we would need to observe as evidence to support those claims, and then the nature of assessment activities giving students the best opportunity to produce that evidence – all within the affordances and constraints of an appropriate mode of assessment delivery. Finally, the data resulting from this domain analysis were initially organized as information instantiated in a specific collection of evidence-centered design objects.

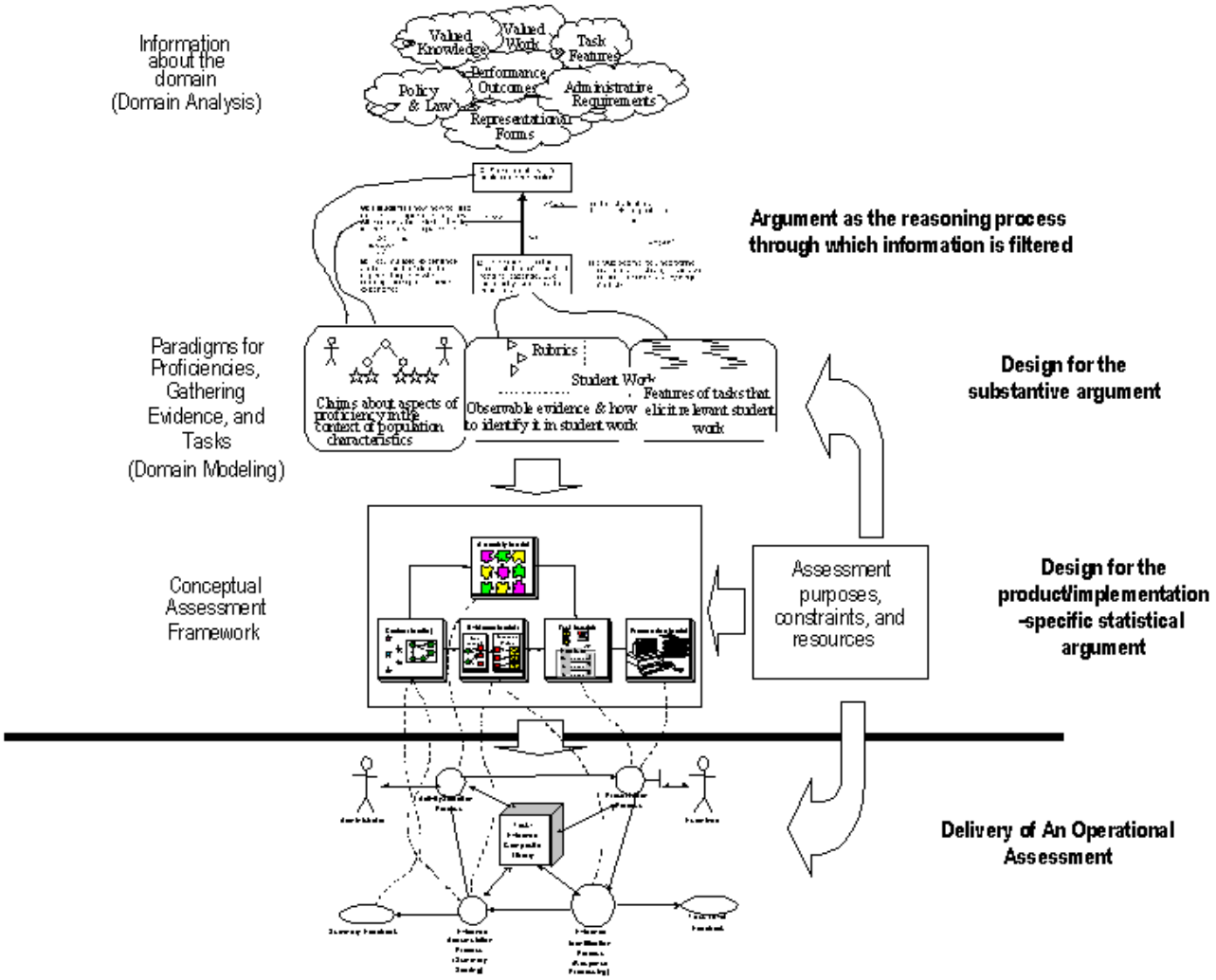

Figure 1. Phases of Assessment Design, with Implications for Assessment Delivery

2.1 The Subject Matter Domain

We chose Biology as the subject matter domain for a number of reasons. First, science education has, in recent years, consistently been the focus of national attention. In consequence, several comprehensive sets of science standards have been developed. Because these standards also reflect common themes related to the kinds of knowledge, skill and abilities deemed important across all science subjects, a science subject provided a good context for our prototype to demonstrate reusability and scalability. From among high school science subjects, biology was chosen because it is the single science course usually taken by most high school students.

2.2 The Domain Experts

A consideration of the assessment purpose was central to the choice of domain expertise that would be needed to support development of the prototype. Because the transitional context (i.e., students moving from high school to college) is an essential part of the assessment's intended use, it was apparent that we needed to engage a set of individuals in the process who could contribute standards-based perspectives of domain understanding at both high school and college levels. In using this variety of expertise, we hoped that the substance developed for the prototype would represent reasonable and consistent expectations at

both sides of the transition: what college professors would want as an acceptable foundation to build on and what high school teachers could actually teach. We were also aware that using challenging, albeit commonly recognized, educational science standards as the basis for the design could potentially result in assessment content, and therefore requirements for student performance, significantly different from current expectations. All the more reason to try to assure that the design process would yield a consensus on 'reasonable and consistent.'

Ann C. H. Kindfield, Ph.D., has extensive background in biology learning research and its application to biology assessment. She has conducted research on the role played by diagrams in learning and reasoning about biological processes and the design of genetics assessments, evaluated genetics learning software, and taught a variety of biology and biology education courses at the college level. She is currently the Biology Education Specialist at Educational Designs Unlimited, an educational consulting firm.

Susan K. Johnson, Ph.D., is a veteran high school biology teacher at Monona Grove High School and a Principle Investigator in the National Center for Improving Student Learning & Achievement in Mathematics & Science at the University of Wisconsin-Madison. She has written and taught numerous introductory and advanced genetics courses for high school students and participated in a number of research and curriculum development project concerning biology learning and teaching.
Scott L. Kight, Ph.D., is an evolutionary biologist at Montclair State University with expertise in science pedagogy. He teaches majors and nonmajors coursework at both undergraduate and graduate levels and is a Leadership Associate of the Center of Pedagogy at MSU.

Gordon L. Mendenhall, Ed.D., is a veteran high school biology teacher at Lawrence North High School in Indianapolis. In addition to his classroom teaching, he has conducted numerous workshops on various aspects of biology education and was a principle contributor to Project Genethics at Ball State University.

Catherine S. Rubin, M.Ed., has experience in science teaching, administration, professional development, curriculum development, assessment design, standards-based teaching and learning, inquiry-based learning and program evaluation. In March 2000 she established EduChange, Inc. to provide educational products and consulting services.

Dirk Vanderklein, Ph.D., is a plant physiological ecologist at Montclair State University with additional training in adult and vocational education. He is a Leadership Associate of the Center of Pedagogy at MSU and has been working with middle school teachers in Montclair, NJ to incorporate inquiry-based teaching into their curriculum.

2.3 The Standards

We used the following collections of science standards as the basis for the prototype design:

National Science Education Standards (1996) National Research Council
Project 2061: Benchmarks for Scientific Literacy (1993) AAAS
New Standards™ Performance Standards (1997) National Center of Education and the Economy and the University of Pittsburgh
Biology Education and Developing Biology Literacy (1993) BSCS

Our domain experts agreed that the sets represented, which are all based on each other to some extent, are commonly accepted in this country at this point in time. They heavily influence the science standards that currently exist or are being developed at the state level.

2.3.1 An Evidence-Centered View of Standards

Most collections of educational standards, in science as well as other domains, attempt to lay out what is important for students to know (e.g., understanding DNA, the concept of an ecosystem) or, to a lesser extent, be able to do (e.g., design and conduct scientific investigation, communicate and defend a scientific argument). Some educational standards also describe activities that afford students an opportunity to demonstrate specific understanding or ability. A very few actually include explicit characterizations of agreed-upon evidence of knowledge, skill and abilities of interest.

Taken all together from an evidence-centered design perspective, educational standards have the following common characteristics:

1) They can be extremely useful in articulating the kinds of knowledge, skill and ability we would like our educational assessments to measure.
2) They frequently omit the essential evidentiary characteristics necessary to the development of performance standards (i.e., how you know when a standard has been met).
3) It is not uncommon for standards descriptions of what knowledge is valued, how you know it when you see it, and/or possible activities for eliciting evidence of knowledge to be bundled in a way that obscures their use in assessment design.
4) The manner in which standards are represented, as discrete pieces of hierarchically organized text, does not succeed in expressing the truly integrated nature of the knowledge, skill and ability it is their intent to foster.

Our response to these characteristics was to set two goals for our standards-based, evidence-centered design:

1) To create an alternative representation of the standards that conveyed their content in a manner better adapted for use in assessment design
2) To demonstrate how, through a process of sorting out and filling in, information available from standards can be aligned with the requirements for assessment design

2.4 The Domain of the Assessment Prototype

We brought our experts together for the purpose of defining the domain of the assessment: the specific biology standards and topics to be targeted by the prototype. We began by establishing a common understanding of the purpose of the assessment prototype.

2.4.1 Purpose

The prototype's primary purpose was to provide a view of what a future culminating (i.e., high stakes, end-of-course) assessment in high school biology might look like when the more complex constructs and student behaviors referenced by standards were targeted. The purpose of the culminating assessment is most naturally thought of as certification of mastery; however, in terms of its underlying models, this purpose is really the same as assessment for selection. We planned to provide scores for this assessment on multiple aspects of proficiency; these would, at least theoretically, be more informative and instructionally useful to the student than a single score (or inaccurately-estimated subscores, measuring skills that are really highly correlated). However, given selection as the purpose of the assessment, there are inherent limitations on feedback to the examinee. The amount of evidence collected is limited by the time constraints typically associated with a selection test; this in turn limits the number and nature of inferences about the state of a student's knowledge that can be supported. Further, all feedback must be provided at the end of the assessment. We understood that these constraints compromised the amount of support for learning we would be able to demonstrate.

What we were confronted with was an all too common problem related to the cardinal principal of test use: assessments that have been designed to optimally fulfill one purpose cannot be validly used to fulfill a

different purpose. When assessment design attempts to fulfill multiple purposes there is usually enough inherent conflict in these purposes that requirements for one will end by taking precedence over all others. We took this as an opportunity to illustrate our solution to this problem.

In order to fulfill our intention to support learning more comprehensively, we decided on an assessment prototype that could be run in either of two modes, distinct yet fundamentally related as to content, evidentiary structure, and delivery systems. The two modes are *culminating*, or high stakes selection or grading at the end of the course, and *interim,* or practice for the culminating assessment that would focus on the same constructs and behaviors and representations for expressing them. In the interim mode we would be able to provide feedback throughout the assessment; the nature of the feedback would go beyond the specifics of any given task performance to both inferences and supplementary material designed to illuminate the constructs common to both modes.

But what stage of the learning process was our interim assessment meant to support? The beginning stages where the assumption is that students need tutoring on the basics of unfamiliar concepts? Much further on in the process where the assumption is that students need practice applying familiar concepts? Some where in between? Applying aspects of evidence-centered design methodology to product planning could yield a whole family of discrete but related assessments targeting progressive stages of achievement (i.e., from tutoring in basics through increasing levels of sophistication, to practice for culminating, to culminating assessment for mastery). Given different purposes, each product would vary with respect to the nature of inference (what we wanted to measure) and feedback (content and timing of information delivered to the student). However, in our project we would be focusing on only one of these possible interim assessments. We also had to be concerned about the extent to which the interim and culminating versions in combination would be able to realize their overarching goal as an effective demonstration of a new kind of high-stakes assessment: one where student achievement in the high-stakes context is supported with instructionally relevant information and learning experiences – in other words, the culminating assessment no longer 'drops in from the sky.'

There was yet a further impact of intended use. Because we were developing for demonstration as opposed to operational use (and on a short schedule), the domain content coverage could not be exhaustive. However, it would have to be sufficiently comprehensive to fulfill the prototype's purposes in a meaningful way. With these considerations in mind, our decision was to focus the interim mode of assessment on practice with application of familiar concepts. Given a conceptual overview of evidence-centered assessment design and this understanding of the prototype's purpose, our domain experts began their work.

2.4.2 Relating Subject Matter Topics and Standards

The initial discussion revolved briefly around the topical structure of biology (genetics, evolution, anatomy, physiology, cellular and molecular basis for biological processes, ecology), but moved rapidly to an expression of concern about the compartmentalization that too frequently seems to characterize science learning. There was immediate consensus that emphasis should be on the larger understanding important for students to acquire about life science in the natural world – the understanding that life is dynamic and interconnected. And further, that the truly valued knowledge in science, in this case biology, is how topical material is both understood in the context of these larger ideas and appropriately related to other topical material also necessary to reveal and illuminate them. These bigger ideas are commonly expressed as major categories in science standards, referred to by our domain experts as **themes**.

The group decided to look at themes first, choose which of them they wanted to focus on, and then pick the biology topics that would best illustrate them. As the conversation progressed, another concern our teachers expressed became evident: that students seem to be lacking in the kinds of abilities required to do scientific inquiry. In consequence, the resulting focus in terms of themes common to science standards narrowed to two. First, there was the **theme of unifying concepts and processes in science** used as a common structure for scientific knowledge. Specific to this theme, our group agreed upon an emphasis on understanding of how specific biological phenomena manifest themselves in across different **levels of organization** (e.g., the molecular, cellular, organism and population levels). A second element of this theme was also chosen as a consequence of the high premium our high school and college teachers set on reasoning skills: **the use of**

models and evidence to reason about and explain biological phenomena. The **theme of scientific inquiry**, of which this kind of reasoning is a key element, is prominent across all science standards. It relates to the nature of scientific inquiry and the abilities, regardless of the level of expertise or professional status of the learner, necessary to do it. It was decided, therefore, to use the inquiry process as the context for the biology learning we hoped to foster.

In moving on to the topics that would best express these themes, the group was of course influenced by their own specific areas of expertise. Given the combination of individual interests and the selected themes, the topics chosen were transmission genetics and microevolution. In particular, combining inquiry with genetics and microevolution seemed especially apt. The ideas that nothing in biology makes sense except in the light of evolution, and nothing in biology is understandable except in the light of genetics, seem to have carried the day.

2.4.3 Representing the Domain of the Assessment through our Domain Experts' Eyes

Once we had agreement from our group about where to concentrate our prototype design, we were faced with the problem of actually trying to represent the standards-based knowledge they valued in a manner that conveyed its integrated nature. Their discussions had illuminated the kinds of knowledge they deemed important for biology students to have in order to master high school biology and to approach biology learning itself. As we have already mentioned, one of the difficulties encountered in interpreting science standards is their representation as discrete, hierarchically organized segments of text (see Figure 1). This is true across all the standards we examined. All of them are consistent in their reference to a common set of high level 'themes' in science; all of them are consistent in elaborating conceptual knowledge within specific topical areas in the sciences. There are, however, no real connections made among these segments of text or, more importantly, among the varieties of knowledge listed. Because the typical representation does not convey how themes in the standards interact with each other or with subject matter topics, we thought that achieving some level of integration by molding this information to conform better with our domain experts' view was an important first step in the development of the assessment's design.

Science Content Standards

Unifying concepts and processes of science (all grades)
- systems, order, and organization
- evidence, models, and explanation
- constancy, change, and measurement
- evolution and equilibrium
- form and function

Science as inquiry (9-12)

Abilities necessary to do scientific inquiry
- identify questions and concepts that guide scientific investigation
- design and conduct scientific investigation
- use technology and mathematics to improve investigation and communication
- form and revise scientific explanations and models using logic and evidence
- recognize and analyze alternative explanations and models
- communicate and defend scientific argument

Understandings about scientific inquiry

Life Science

Molecular Basis of Heredity

Theories/Models: chromosome theory of inheritance, chromosome mapping, semi-conservative hypothesis, base-pair complementarity, chi-square model, Punnett squares

Figure 1. A typical textual representation of science standards

In the end, the job of evidence centered assessment design is, through an iterative process, to sift and mold information, regardless of form or content, into the essential operational components of an assessment: student, evidence and task models. The relationship between form and function in assessment models is just as compelling as, for example, the relationship between the double helix and the replication of genetic material. Therefore it was natural for us to continue our domain analysis by attempting to mold the information gleaned from our experts into forms that more closely approximated those of operational models. Operational models consist of clusters of variables organized into structures. The variables themselves have both semantic and quantitative meaning, as do the relationships that define their internal structure and their inter-model structure. To support this phase of design process, we took each part of the standards chosen by the group, re-represented it, and finally integrated the pieces into an alternative standards-based view of the domain of the prototype. The elements of this view (see Figure 2) resulted from a cross-standards analysis in which we focused on common themes as well as elements from unifying concepts and processes. In addition, we used concept maps to guide our analysis of topical content.

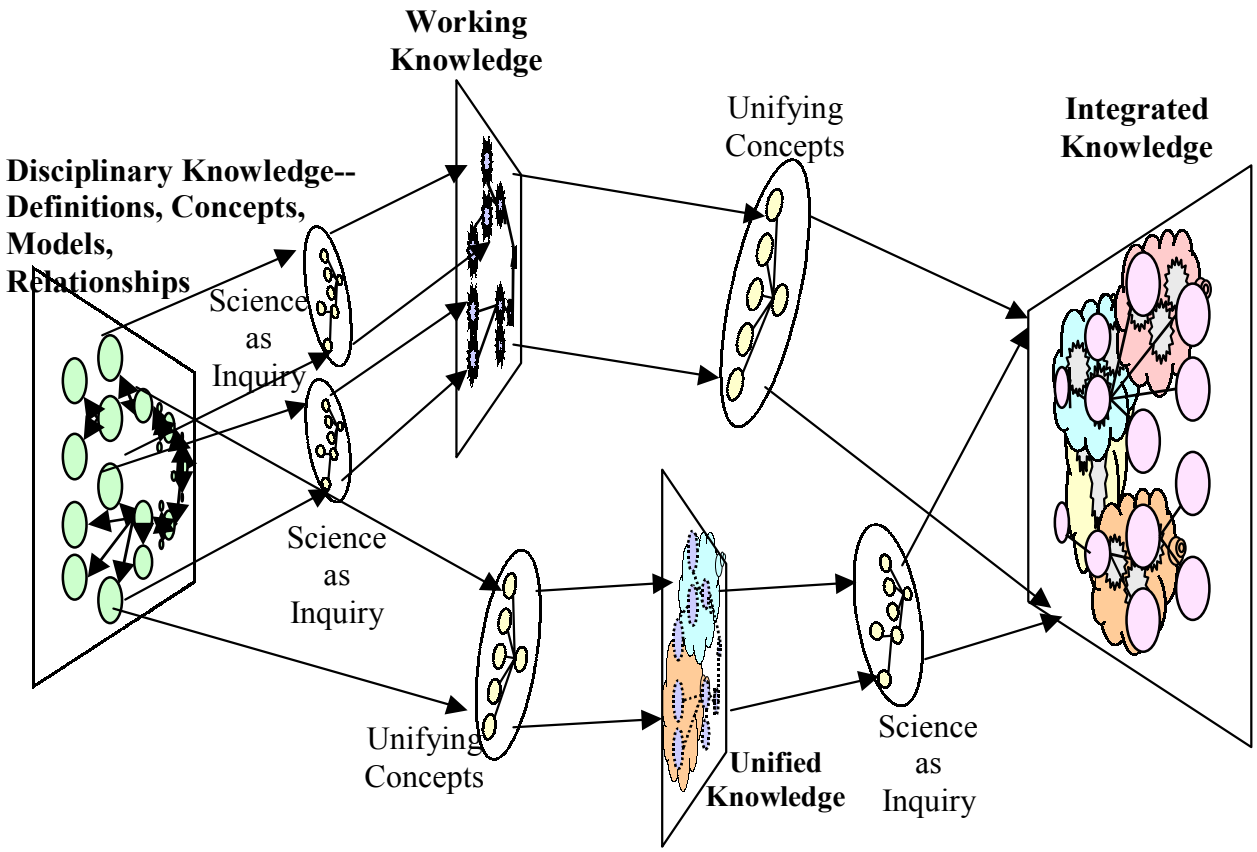

Figure 2. Representation of a Standards-based Domain for Assessment Design

Let's step through the diagram from left to right. Disciplinary knowledge is the simplest form that biology knowledge can take. At the far left we see, in perspective, the same knowledge of the definitions, models, and relationships in transmission genetics and microevolution that was represented in Figure 1. Moving to the right, two different things happen. Disciplinary knowledge can be extended in one way by understanding how to use it as the substance of scientific inquiry. In this transformation (which produces what we call Working Knowledge), understanding of definitions, concepts, models, and relationships has synthesized and developed to the point that students actually know how to use it meaningfully in the context of inquiry – whether it's being able to explain a particular phenomenon in terms of one or more

underlying models, or investigating the plausibility of a given aspect of an explanatory model. As the diagram shows, disciplinary knowledge can also be extended by understanding how it relates to a particular unifying concept or process – for example, looking at cells through the unifying concept lens of form and function, helps us understand how cellular structures facilitate cellular processes. The view through any one of the unifying concepts gives structure and real explanatory power to myriad pieces of disciplinary knowledge. Different unifying concepts help us organize disciplinary knowledge in new ways to answer new questions. The lens of a unifying concept can also be brought to bear on the Working Knowledge we construct. We can take a working model of the cell and its various processes and, by viewing it through a conceptual lens that emphasizes systems, order, and organization, organize and acquire knowledge that helps us understand how the work of the cell helps accomplish the work of the organism. Finally, at the far right, students develop fully integrated understanding which allows them to use models, evidence and explanations from different topical areas and different levels of organization to identify and answer increasingly bigger questions– knowledge that is both vertically and horizontally integrated.

Since the type of information supported most strongly in standards is description of knowledge, skills and abilities, the view in Figure 2 represents the first step in developing our prototype's student model – which defines what the measurement model's inferential machinery will make inferences about. Further, one can view this figure as a **generalized schema** for representing the knowledge, skills and abilities valued in science standards. (As we continue through the design process, the utility of this approach will become evident.) The lack of cross-standard descriptions of evidentiary requirements, or how you know when a standard has been met, is both a strength and a weakness. It is a strength because, given local constraints, it allows for more local control of what students are expected to produce. However it is a weakness for the same reason, because there is no explicit consensus about what it means to meet a standard. This has important consequences for the rest of the domain analysis. In the absence of this information our experts had to develop their own evidentiary criteria. Only then would we be able to design the tasks that could elicit the behaviors our experts were looking for.

2.5 Focus on Defining Claims, Evidence and Tasks

Once we had represented the knowledge valued in the domain of the assessment at the highest level, we began the process of refining our focus. Using evidence-centered design methodology, we concentrated on understanding the claims our experts thought would be important to make about students who participated in this kind of assessment. From there we moved on to identifying the kinds of behaviors and features of work (evidence) teachers would expect to see their students produce at different levels of proficiency, and finally to identifying those elements of various performance situations that would make it possible to elicit the looked-for work and behaviors.

2.5.1 Claims

Claims are the fundamental building blocks of assessment design. The significance of claims lies in their power to connect the purpose and audience of the assessment with its inferential requirements. Claims define what you want to be able to **say about a student** as the consequence of assessment. Since we can never observe the true state of a student's knowledge, skill or ability, claims are really inferences. Regardless of the form in which the claim is made to the audience for the assessment – whether one or more simple statements on a report, placement on a proficiency scale, or via an interpretive guide to score meaning -- one thing is sure: valid claims, or inferences, are those that are supported by evidence. In order to support a claim with evidence, it must some how be represented in the student model. It follows then that defining claims moves us along toward two design objectives: developing the semantic meaning of an operational student model, and laying out the characteristics of our assessment reporting.

There are two ways we can assure that any given claim we want to make is supported by evidence. The first way is to represent it with its own variable in the design – even if, for whatever reasons, this variable will not be included in the operational student model. This allows for the **definition of evidence related specifically to that claim**. Evidence related to the claim contributes to an accumulation that supports a more general inference, but by formally modeling the claim with its evidentiary requirements as part of design we assure that the more general inference can be validly interpreted with the meaning intended. For

example, a single variable (e.g., theta) in a Student Model that is used to accumulate evidence related to verbal proficiency can be interpreted in reporting to support several specific claims if the evidence for these claims has been explicitly modeled during design. What exactly is meant by verbal proficiency is defined by what you can say about the examinee **based on the evidence** you have collected. It's easy to imagine the number and range of different claims one could make about verbal proficiency – from simple inferences about vocabulary use to much more complex claims about comprehension – what you can say depends entirely on the evidence you collect. The evidence you collect depends on how you have elaborated knowledge, skills and abilities in the construct of interest. The second way of assuring validity of a claim is to operationalize the claim with its own variable in the student model. This allows for **the accumulation of evidence related specifically to that claim**. (It goes without saying that this evidence needs to be modeled in the design.)

The standards-based nature of our prototype meant that we would have to articulate our claims in terms of the disciplinary and thematic knowledge illustrated in the domain representation. Within these we would have to integrate (1) the disciplinary topics of interest, exemplified in Figure 1, and (2) targeted components of inquiry and the unifying concepts, depicted in Figure 2.

Looking at these representations it's clear they are a source of many aspects of knowledge we might want to measure. The number of claims we could potentially articulate (the number of inferences we could make) from a representation like this is infinite – in fact, a universe. For example, there are many more content subareas in secondary biology, and there are additional Unifying Themes and aspects of Science as Inquiry that we have not discussed. Different assessments might focus on different claims, at different levels, different grain-sizes, and different disciplinary areas and subareas, according to the purpose of the assessment. The claims we have discussed are consistent with our purposes--a culminating and interim assessment in particular subareas. Other sets of claims would ground a coached practice system, for example, or a multiple-choice test of terminology. Figure 3 contains a few of the claims formulated by our experts for Biomass.

Some Claims about Disciplinary Knowledge

Understands the entities, events, and outcomes constituting the Mendelian model
Understands the entities, events, and outcomes constituting the sexual life cycle
a. Understands the entities, events, and outcomes constituting meiosis
b. Understands the entities, events, and outcomes constituting the chromosome model
c. Understands the entities, events, and outcomes constituting ploidy

Understands the entities, events, and outcomes constituting natural selection
Understands the entities, events, and outcomes constituting genetic drift

Understands the elements of an experimental scientific investigation in the context of transmission genetics.

Understands the elements of a naturalistic scientific investigation in the context of mechanisms of evolution

Some Claims about Working Knowledge

Given complete data, can
• Use data to explore natural phenomena at the population level(s) across mechanisms of evolution and across time...
a. When exploring natural phenomena entails identifying and reasoning through models for the purpose of explanation and prediction
b. Where models involved are natural selection and genetic drift

Figure 3. Some Biomass Claims

2.5.2 The Proficiency Paradigm

Once we had identified the claims that the prototype would actually support, we began refining the representation. Creating a sketch of the Student Model is the next step in formalizing the semantic meaning of the relationships among aspects of knowledge, skill and ability of interest to us. In this sketch, called a

Proficiency Paradigm, there is still no description of the statistical properties of the relationships. Figure 4 is an example of a Proficiency Paradigm that lays our construct for both the interim and culminating assessments. The relationships defined at the highest levels of disciplinary, working, and integrated knowledge indicate the necessity for a fairly complex student model, one that more closely approximates realistic knowledge interdependencies in the domain (Kindfield, 1999, Stewart & Hafner, 1991). While any bit of evidence is designed to bear on one or more given variables(s), the interconnectedness of the aspects of proficiency means that same bit of evidence will have indirect impact on the rest of the model as well. Therefore, if evidence bearing on the claim that a student understands the entities, events and outcomes constituting the Mendelian model (represented by the aspect of proficiency labeled DK Mendelian Model) is absorbed, this evidence will also change, in some way, our belief about the student's ability to use the Mendelian model in the context of inquiry or one of the other standards-based themes. In this example in particular, direct evidence of lack of disciplinary knowledge about the Mendelian model leads us to expect the student will have trouble in situations that call for applying such knowledge in the course of an investigation.

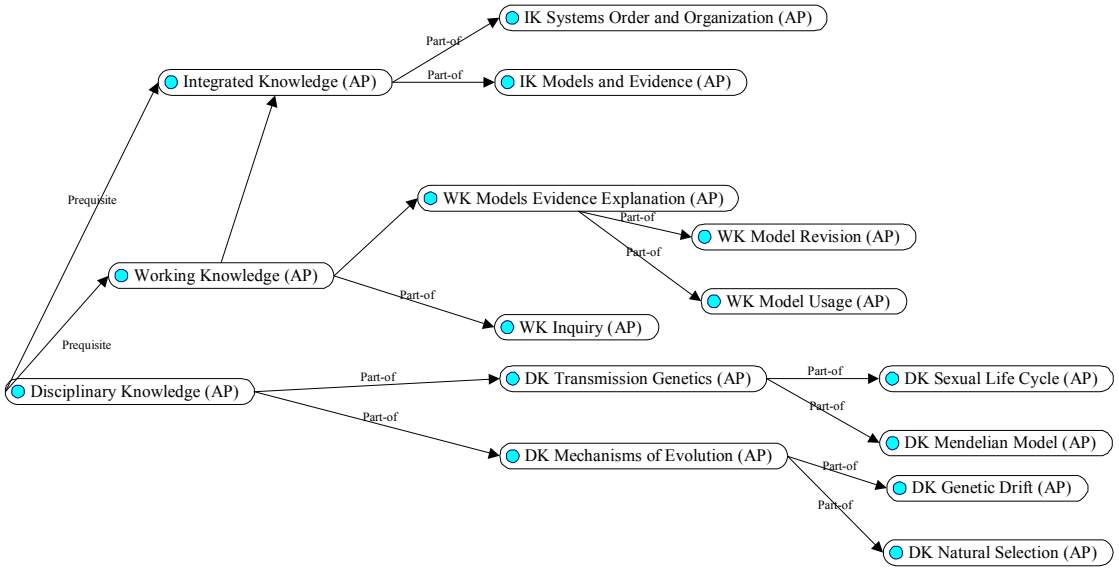

Figure 4. A Biomass Proficiency Paradigm, at a grainsize meant to support interim assessment use

Figure 5 depicts the Proficiency Paradigm for the culminating assessment only, now with claims included. The aspects of proficiency are those from Figure 4, except those at the finest level of detail are omitted. It is not anticipated that there would be enough information about them to report on separately. Note that the claims for Working Knowledge have, as parents, aspects of both disciplinary knowledge and working knowledge proficiencies, for both are presumed to be necessary to say that a student has actionable knowledge with respect to given content. Analogously, the proficiency parents of the Integrated Knowledge claim are the working knowledge proficiency and the required aspects of disciplinary knowledge.

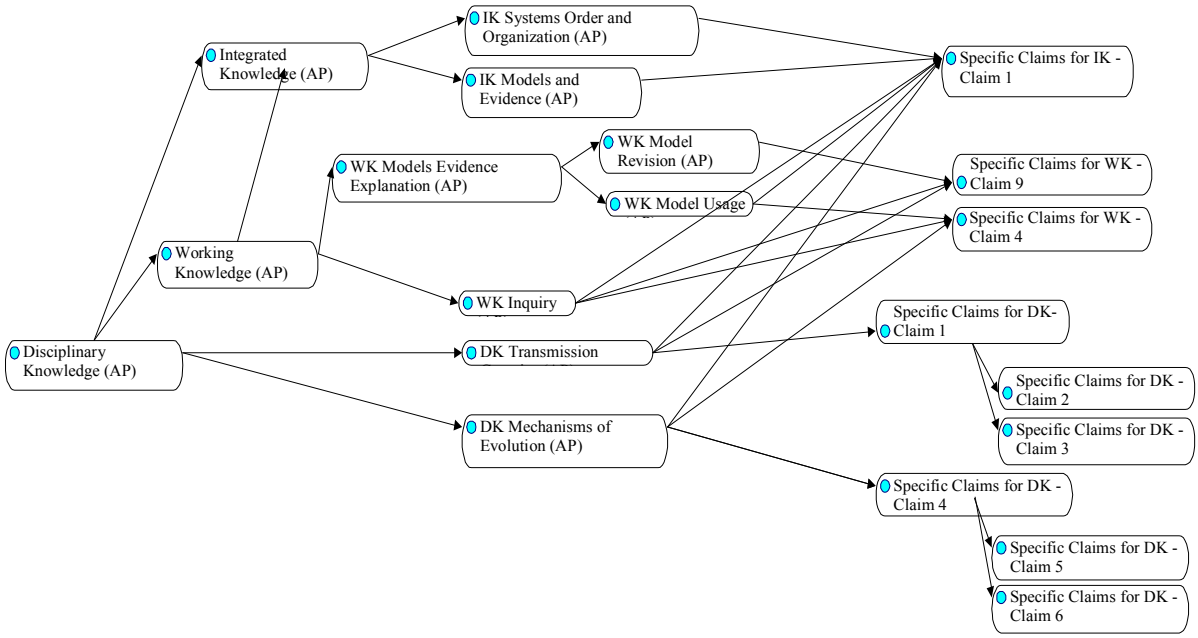

Figure 5. A Biomass Proficiency Paradigm with Claims

In this movement from domain analysis to domain modeling, the structures we create also prompt the collection of additional information; in this case, a description of one or more states for each of the nascent student model variables corresponding to the levels of proficiency we want to use in our score reporting.

By comparing Figures 2.4 and 2.5 we can see that the paradigm without the claims appears quite a bit more abstract, as indeed it is. The claims give each aspect meaning – but not an absolute meaning, only one that is appropriate for our particular assessment product. The claims guide us in defining the evidence we need to collect. This has important implications for reuse of design elements. We can easily change the claims associated with one or more of the aspects; but as soon as we do, our evidentiary requirements change as does the validity of the assessment for its stated purpose.

The Proficiency Paradigm as sketched is also scalable. If we wish to articulate new claims, we can modify the sketch to add, for example, new components of the same themes; we can add new themes; we can further decompose or use additional biology disciplinary knowledge. We can even, still keeping within the universe that a standards-based domain representation provides, add a different body of science content.

2.5.3 Evidence

Once we had established what we wanted to measure with some descriptive clarity, we set about defining the body of evidence that would be needed to support our claims. In assessment, evidence comes to us from what we can observe about a student's behavior or work. As mentioned earlier, the science standards do not provide much information about commonly accepted observations that tell us whether or not a student is performing at a particular level of proficiency. At best, given a specific activity, some standards identify desirable qualities of work products produced in response to that activity. Therefore we turned to our experts for guidance.

At this point the ECD process focused on each individual claim to be supported by the prototype. Evidence, or a set of observations, related to each claim was considered independent of any specific task, although some observations were naturally more suggestive of generic activities a student might be asked to undertake. Since most disciplinary knowledge would be contextualized as part of working or integrated

knowledge, primary attention was given to thinking about and specifying observations at the Working and Integrated knowledge levels. Also, evidence of much disciplinary knowledge was straightforward and its definition could be aligned with the concept maps the experts had already developed for the demonstration.

As an example, Figure 6 contains some observations that our experts thought would be needed to support the claim that a student could *design and conduct a scientific investigation*.

Recognition of necessity to produce more data
    Efficacious specification of appropriate scientific methodology(s)
    Association of anomalous data with relevant element(s) of relevant model(s)
        Impasse specified in terms of data/model (what's the mismatch)
Accuracy of Model changes
Adequacy of Model testing
    Efficacious specification of appropriate scientific methodology(s)
Identification of outcomes of model testing that bear on current hypothesis
        (confirm/disconfirm)

The following observation bears on pre-requisite disciplinary knowledge
    Recognition and/or use of entities, events and outcomes of Mendelian model

Figure 6. Some observations related to scientific investigation

2.5.3.1 Evidence Paradigms

We were now ready for the next step: sketching potential evidence models, known as Evidence Paradigms. As with the Proficiency Paradigms, Evidence Paradigms are descriptive; but they contain important information in addition to semantic meaning. Specifically, they begin to impose a certain structuring, or relationships, among particular kinds of information that will subsequently inform the construction of Evidence Models. Evidence models are the most complex design objects because they form the bridge between what we want to measure and the tasks we develop to elicit evidence. This means they have to contain enough information to form coherent links in both directions. As a step in this direction, developing Evidence Paradigms accomplishes the following:

- Transformation of descriptions of observations into a collection of discrete observable variables
- Definition of one or more states for each observable variable
- Definition of knowledge representations as the basis of work product design.
- Definition of rules (or rubrics) for evaluating work products to extract their observable features and produce specific values for them.
- Initial specification of impact of observables on the student model; that is, qulaitative, but not yet quantitative, information about which aspects of proficiency are reflected by which aspects of performance.

2.5.3.1.1 Observables

In the powerful paradigm of standardized testing as it is usually implemented, we have come to assume that what we can observe about a student's response is that it is either right or wrong, or somewhere in between. In fact, while right or wrong may constitute the 'bottom line' for some assessment purposes, it does not tell us anything about the nature of the evidence the student is providing. This is critical to capture since there

is no way to validly reuse tasks unless we know what kind of evidence they provide. In trying to define the nature of the evidence, we also define the scope of our evidence-collecting job.

Consequently, an essential part of domain modeling is to formalize and preserve descriptions of evidence with the creation of proto evidence model variables (which have no statistical attributes, only qualitative attributes). Figure 7 contains descriptions of a small subset of the collection of observables we defined for the prototype. Each of these proto evidence model variables is derived from the evidentiary descriptions provided by our experts. The result is that the observables represent *classes* of observations because they are still task independent. This makes these observables reusable across tasks. The value, or state, of any evidence model variable is set as a result of evaluation of student work. Associated with each observable in the figure is a description of what the essential features of that specific bit of evidence would look like at three different levels of proficiency, the number of states we decided to work with for most of the Biomass observable variables.

Mendelian Model Representation
Description:
Representation of various aspects of the Mendelian Model
using recognized symbolic language of the domain.
Possible Values:
All aspects of Mendel's Model represented using symbolic forms
Some aspects of Mendel's Model represented using symbolic forms
Symbolic forms not used to represent aspects of Mendel's Model

Efficacious Methodology
Description:
Efficacious specification of appropriate scientific methodology(s).
This includes appropriate data generation and data analysis
methodology as well as hypothesis generation within the
framework of the hypothetico-deductive method.
Possible Values:
Models revised/tested w/efficacious scientific methodology
Models revised/tested w/somewhat efficacious scientific methodology
Models not revised/tested w/efficacious scientific methodology

Data/Model Relationships
Description:
Relationship(s) between patterns of data
and particular models
Possible Values:
Data and model(s) related appropriately
Data and model(s) related somewhat appropriately
Data and model(s) not related

Anomalous Data/Model Connection
Description:
Association of anomalous data with relevant element(s) of relevant model(s)
Possible Values:
Anomalous data associated w/model element(s) requiring revision
Some anomalous data associated w/model element(s) requiring revision
Anomalous data not associated w/model element(s) requiring revision

Figure 7. Some Biomass Observables

A note in passing about "efficacious methodology": "Efficacious" means both effective and efficient, and describes ways that experts in a domain often attack problems and use tools of the trade. A good example is Kindfield's (1999) research on the diagrams that subjects spontaneously drew to solve a series of

problems in cell division. Those least skilled often could not find a suitable representation to solve the problem. Students with more experience drew very thorough diagrams, and usually solved the problem. Somewhat surprisingly at first, the diagrams of experts were cruder than thos of students; they included fewer details, and looked less like those in textbooks. A closer look revealed, however, that their diagrams included only the features that were most directly relevant to the problem at hand, and lacked details and relationships the novices depicted, but were immaterial to the task. In a word, the experts' use of diagrams was efficacious.

2.5.3.1.2 Observables and Knowledge Representations

Once we had defined our collection of observations, we had to start developing the bi-directional linkages between them and our Proficiency Paradigms on the one hand and some idea of tasks on the other hand. This is true because the value and impact of any observation on our belief about a student's proficiency (eventually driven by the structural and statistical properties of the student model) is shaped by the nature of the work the student produces, and the work that a student produces results from performance of a specific task. For example, if we wanted evidence of a student's writing proficiency we could observe the logical organization in an essay. However if that essay were produced as the product of a collaborative group, the observation (and any others that could be made, such as sentence structure or vocabulary usage) constitute far less convincing evidence of individual proficiency than observations made of work produced by the student *alone* – an essential feature of the task. Or, as another example, observing logical progression in a mathematical explanation may provide more powerful evidence if the work product is a proof rather than simply the steps in a problem solution. Therefore, while design work to this point had been iterative but still moving along one path, the development of evidence models necessitated looking in two directions at once, working jointly with multiple design elements of proficency, evidence, and task paradigms.

We began with the fact that observables can only be evaluated (i.e., assigned one of their possible values) by being assigned to specific work a student has produced. Our goal was to define a collection of types of student work that would be likely vehicles for the data needed to make our observations. To this end we drew on important information we had gathered from our experts in the domain analysis: knowledge representations.

Once they had decided on the topics within biology they wanted to target, we had turned their attention to descriptions of how information about these topics would typically be communicated within a biology learning community. In mathematics, for example, information is communicated via collections of symbols used to describe number systems, laws, and operations; relationships may be described using symbols or graphs. In chemistry, there is a whole different 'language' (or representation) used to communicate information: the symbols of the Periodic Table of Elements and rules for expressing compounds in their terms. Physics has its formulae, history its maps and timelines. An analysis of any discipline produces a multiplicity of representations that are valued as vehicles for conveying information in that domain. Even with common representations such as text, conventions within a community act to customize them, as with the elements of and standards for a research report or a piece of expository writing. Different kinds of knowledge within a given domain are represented in different ways. In algebra we are comfortable with using the symbol $\boldsymbol{X}$ to represent a discrete piece of information we think of as the unknown. We are also comfortable with expressions like "a + b = b + a," the representation for one of a collection of rules constituting the model for operations within that domain.

In biology, specifically within transmission genetics and microevolution, there too are conventional forms for conveying information: Punnett Squares, phenotypic distributions, allele symbols, pedigree and chromosome diagrams, and population tables just to name a few. In assessment design, identifying the salient knowledge representations for a given domain helps us think about how information is conveyed both to and from the student. When information is coming from the student, we think about representations the student can use to create a work product. When information is being conveyed to a student, we think about the representations we need to use in creating material presented to the student. As part of their work on knowledge representations, our experts identified different ways of representing Mendel's model (also known as a mode of inheritance). Figures 2.8 and 2.9 present two different representations valued for

conveying information about mode of inheritance. Figure 10 shows the same information represented textually. It is an interesting exercise to compare these in terms of evidentiary potential.

Chromosome Type	Alleles	Genotype/Phenotype	Dominance Relationships
A_n	Ag-1	Ag-1Ag-1/agouti	Ag-1 co-dominant with respect to Ag-2
	Ag-2	Ag-1Ag-2/agouti-tan	Ag-2 co-dominant with respect to Ag-1
		Ag-2Ag-2/black-tan	

Figure 8. A Traditional Representation of a Mode of Inheritance (the Mendelian Model)

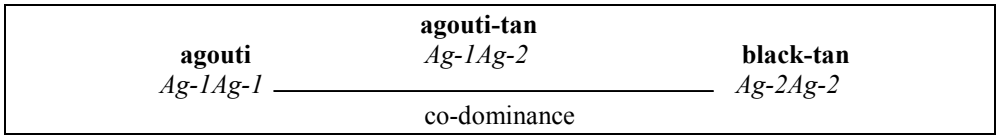

Figure 9. An Alternate Representation of a Mode of Inheritance (the Mendelian Model)

The gene for coat color is an autosome.
There are two alleles for this gene in the population
When the two alleles are in the same individual, they both show up in that individual's coat color.

Figure 10. A Textual Representation of a Mode of Inheritance (the Mendelian Model)

In general, the knowledge representations used in transmission genetics consisted of the kinds of symbols shown above as well as tabular data, an example of which is shown in Figure 11. Figure 12 is an example of the tabular representation of data typically emphasized in microevolution.

Cross	Offspring		
Agouti f X Agouti m	11 Agouti (6 f and 5 m)		
Agouti-tan f X Agouti-tan m	3 Agouti (2 f and 1 m)	7 Agouti-tan (3 f and 4 m)	2 Black-tan (1 f and 1 m)
Black-tan f X Black-tan m			10 Black-tan (5 f and 5 m)

Figure 11. A Common Representation of Population Data in Transmission Genetics

	long tail, no collar, & purple head	long tail, collar, & purple head	short tail, no collar, & purple head	short tail, collar, & purple head	Total in sample
t_0	31 (15f: 16m)	9 (6f:3m)	29 (14f:15m)	11 (4f:7m)	80 (39f:41m)

Figure 12. A Common Representation of Population Data in Microevolution

Knowledge representations are primary links between evidence and tasks. Therefore, we consider them to be a good first step in task design. It should be evident from these examples that it's possible to go quite a long way in defining evidence before having to grapple with the idea of a specific way of getting it. In fact, we would argue that thinking about the relationship between what you want to observe and the way knowledge is conveyed within a domain absent the idea of a specific type of task is a good way to broaden ideas of what can and should be considered as evidence and ways of getting it. Already, given these kinds of representations, it was becoming clear that the data-driven nature of working in these areas of biology would result in tasks that emphasized the manipulation and interpretation of data – tasks that are quite different from those traditionally seen in standardized Biology assessments. If we look back at the standards-based domain representation (Figure 2) we can anticipate that as we move from left to right the number of different knowledge representations necessary for conveying information increases; one could reasonably expect that facility with multiple knowledge representations would be essential to evidence in support of Integrated Knowledge claims. The impact of the science standards was already becoming apparent.

Because we can associate information in these representations with specific aspects of evidence (observables), having students work with or create them in some way will allow us to produce evaluations for such observations as Mendelian Model Representation (and others) of that bear on our DK Mendelian Model aspect of proficiency. True enough, this may be evidence of fairly rudimentary knowledge, and, as such, is weighed accordingly. More compelling or sophisticated data for observations may be found in other knowledge representations (such as phenotypic distributions).

Our experts also identified a number of ways of communicating about investigative methodology in the contexts of transmission genetics and microevolution. At a high level, there are the steps in the hypothetico-deductive framework (as shown in Figures 2.13 and 2.14). At a lower level, there are the rules governing the selection of test populations and individuals within them. Having students work with these

rules within this framework will give us information we can evaluate for observables like Efficacious Methodology. Again, depending on how these representations are realized within a task, and what the student is asked to do with them, they will provide differentially valuable observable features. Note that the generic nature of each observable allows for it to be connected with any number of different representations -- each of which may vary in the completeness, depth or accuracy of observable features – making up potential work products. Again, the observables and knowledge representations were pointing us in the direction of novel kinds of tasks.

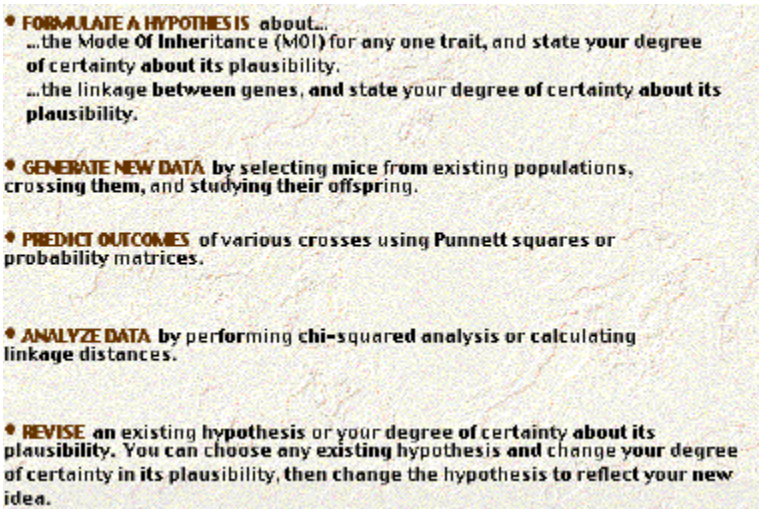

Figure 13. A Representation of Investigative Methodology

Methodology	Associated Knowledge Representation
FORMULATE H_0	Hypothesis expressed in standard form or alternative form
GENERATE DATA	Population Summary Cross Table; Cross Choice Table
ANALYZE DATA	Hypothesis expressed in standard form or alternative form; Population Summary Cross Table; chi-sq table
ACCEPT H_0	Hypothesis expressed in standard form or alternative form; Population Summary Cross/H_0 Connections Table

Figure 14. Connecting Knowledge Representations

In sum, before one can move forward with further refinement to determine the specific evidentiary 'yield' in any given situation, connecting aspects of evidence (observables) with the representations constituting student work is necessary and important for the following reasons. First, it makes a direct link between some objective (non-task dependent) definition of evidence and how information is communicated within the domain. Second, considering knowledge representations as the sine qua non of tasks provides essential guidance for task design. Finally, explicit consideration of knowledge representations as central to task design sharpens focus when consensus on evidentiary requirements is sought, and reaching such consensus increases the validity of the assessment. For the next step, we moved to tasks.

2.5.4 Tasks

In order to continue fulfilling the knowledge requirements to proceed with our evidence paradigms, we need to refine the evaluation of the observables that evidence the proficiencies. In particular, we need to think about specific task situations and kinds of tasks that produce particular instances of knowledge representations as work products.

2.5.4.1 Task Organization

By doing the work described above, the challenge of task design can be phrased as follows: how to get students to interact with and/or produce the collection of knowledge representations we need to evaluate observables. Before we addressed any particular task, we needed tothink through the characteristics of the collection of tasks as a whole that we would implement for the prototype. Table 1 summarized important implications from our work up to this point.

Table 1: Design Facts and Implications for Task Design

Design Fact	Implications for task design
The purpose of the prototype mandated: • a high stakes assessment • a learning assessment to support high-stakes achievment	1. At least one collection of tasks for each mode 2. Common use of representations across collections
Student model sketch designed to support three major claims, at least two of which are related to scientific investigation	Tasks within a collection would be organized by focal claim
Evidentary requirements to support each claim resulted in a collection of observables that was large and diverse within a collection.	1. Multiple tasks (items) would be required be required for each claim 2. Observables would be distributed over task/claim groupings within a collection

Figure 15 plays out these implications using a single focal claim as an example. There are two overall collections of tasks, both focusing on the same aspects of skill and knowledge. The learning tasks are organized into a scenario, and stream across the figure. The first five segments are drawn out more fully, indicating their order, the feedback at the end of segment, and the way that culminating assessment tasks can be developed from one or two segments. The rest of the segments from the interim scenario are listed along the bottom of the figure. Together, the seventeen segments of the interim task constitute phases of an in-depth investigation that students would work through, perhaps over the course of several days, as a part of their instruction. Students might work on them individually, in groups, or together as a class.

The culminating tasks (grouped into testlets) appear above them in the figure, each growing directly out of its related learning segment. They are more focused and less extended investigations, as would likely be required in the setting of time constraints and individual work. Common observables bearing on the indicated aspects of proficiency are distributed across the learning and culminating tasks. Note that this relationship between learning and culminating tasks allows students to become familiar with interfaces, knowledge representations, and expectations for evaluation during the course of study, so that these necessary components of complex tasks will not 'drop in from the sky' on them in the culminating assessment.

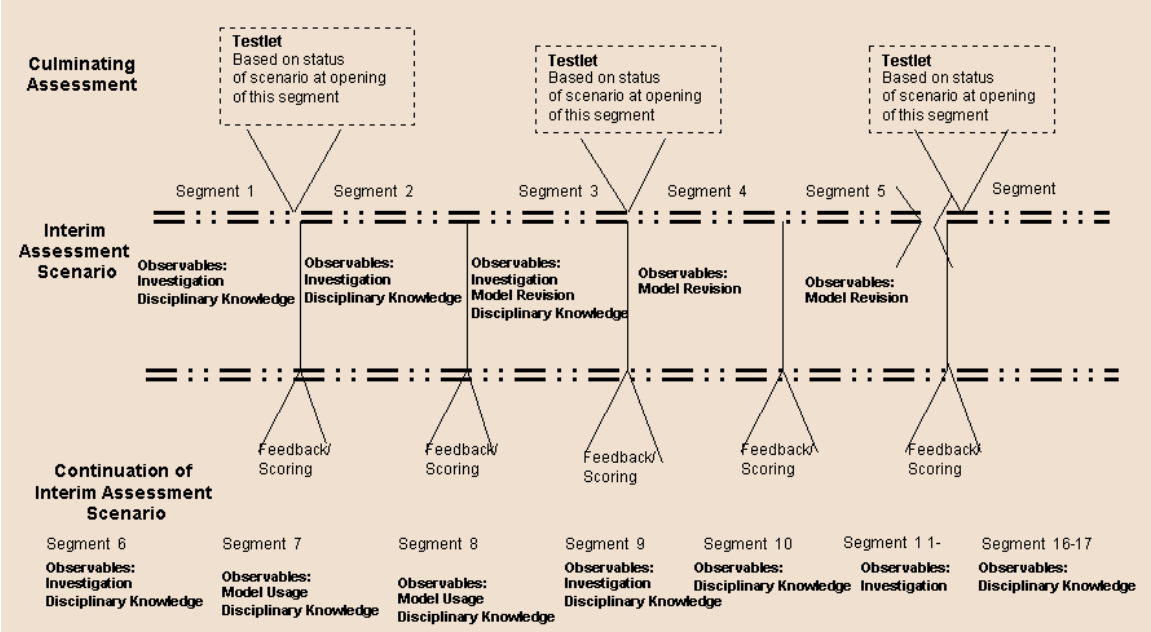

Figure 15. Organization and Relationship of Culminating and Interim Assessment . Focus is a Working Knowledge Claim: Conducting Investigation and Revising Models in Transmission Genetics Tasks

2.5.4.2 Using Hierarchical Task Paradigms

Figure 16 is a hierarchically schematized view of task organization for the Working Knowledge Claim represented as the organizing theme in Figure 15. Each lower level down represents an an increase in task specification; each level serves as a container for the next lower level.

Using this view, we can give the following task-oriented description of our prototype: Our collection of learning task models consists of one task model for each claim; the claim task model consists of one task model for each scenario; the scenario task model consists of one task model for each segment; the segment task model consists of one task model for each task. Each task contains the knowledge representations to be presented and to be collected as work products. Features set at each level further help to specify, or constrain, choices of both features and feature values at the next lower level. The collection of culminating tasks was shaped in a similar fashion; however, there is an explicit connection between the task model for a learning segment and the task model for a culminating testlet. This connection is akin to a pedigree in that the culminating tasks were developed as a reflection of segments presented as part of the learning experience. This was to ensure that the culminating assessment did not appear to 'drop in from the sky.' The testlet could be a copy of a segment with changes to surface features, or the testlet could be derived by compressing multiple segments. Either way, students would be able to recognize both the representations and the activities in culminating assessment as relevant to their learning experiences.

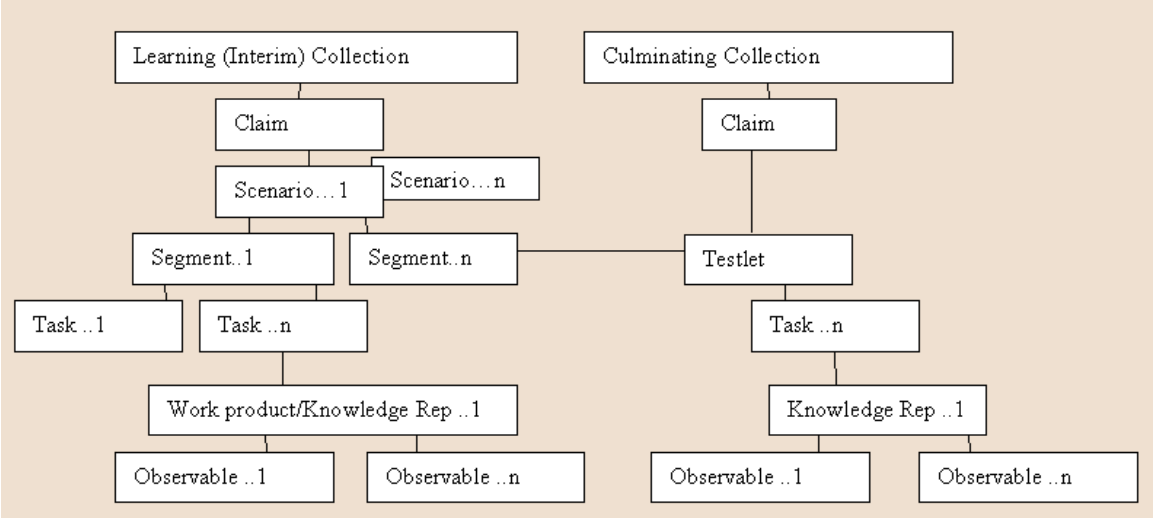

Figure 16. A hierarchical organization of task models

At the highest level, we created task paradigms. These are models, or schemas, for specifying task content unrestricted by implementation or delivery requirements. The purpose of these task pardigms was to begin to provide support for subsequent development of specific tasks by first specifying the salient features defining classes or families of tasks. At the highest level these features described the nature of purpose, domain, audience, platform, and feedback options. At the Claim level, features specifying the type of knowledge, domain topics, nature and number of models were described. At the next lower level, we started shaping the specific tasks more directly by specifying the general form (e.g., scenario) in which individual tasks would appear, the nature of help and guidance, the level of organization for the content (e.g., Disciplinary, Working, Integrated), any problem constraints (e.g., number of traits) , the type of activity to be carried out (e.g., field investigation), constraints on that activity (e.g., population sizes, nature of the 'field') and additional content specification (e.g., organism). When we got to the level of sketch that described a segment, we were essentially decomposing previously specified information into smaller grainsizes. At the individual task level we dealt with specific features of knowledge representations that are presented as problem, reference, or response data in instances of particular task forms (item types).

Before we could complete our evidence paradigms, task paradigms had to be completed to the point that the knowledge representations the student would need to use in producing a work product were fairly well specified. However, the exact manner in which the student would produce the work product did not yet need to be specified. For example, in the kind of task using the knowledge representation from Figure 9 (which was realized as part of the first segment of the scenario organized around Working Knowledge claim 9) it was necessary in the task paradigm only to specify that the problem was to provide a mode of inheritance and also to specify the symbols to be presented. At this level it is not necessary to know how the student will interact with the symbols to identify the mode of inheritance (e.g., drag/drop, text entry) – only which ones will be presented. Specification of these symbols is, of course, guided by the higher level task paradigm which describe several features which act to constrain the content and requirements of the investigation (such as number of genes).

Within any task paradigm, task features can be assigned to any of several categories (such as Size/Complexity, Help/Guidance, Setting) and can also be given any of several potential roles (Difficulty, Evidentiary Focus, Task Selection, Realism) – all in order to more completely describe how that feature is intended to play out within the sketch. (see Mislevy, Steinberg, & Almond, 2002, on the roles of task model variables.) As the tasks move from design through implementation, task model features are 'absorbed' as they are used to realize elements of the task (e.g., the materials presented to the student). That is, the values they take on for a given task are embodied as characteristics of the implemented task. In later stages of design, these task paradigms will be used to develop tasks models which, in addition to

specification of content, include the features a task will need to move into the operational environment. These are the features that are used by the various assessment delivery processes. This is discussed in Implications for the Assessment Delivery Processes further on in this section.

When we had finished this stage of our design work, then, we had described task paradigms (schemas) for both a culminating collection and a learning collections of tasks. Each collection consisted of either testlets or scenarios, each of which was designed around a specific claim. The same claims were used as the basis for both learning and culminating collections. As mentioned before, we used our understanding of what we would have to observe and the nature of student work that would afford us an opportunity for those observations to design the culminating testlets as outgrowths of one or more learning segments.

At this stage the learning collection consisted of the description of two different learning scenarios. The first was designed as an **experimental** investigation in transmission genetics, where the emphasis was on use and revision of the Mendelian Model to come up with a mode of inheritance for coat color in agouti mice (Working Knowledge Claim 9). The problem was constrained to a single trait controlled by three different forms (alleles) of a single gene. The investigation activity was constrained by being situated in the classroom with the teacher present to provide information and guidance.

The second learning scenario was designed as an **observational** investigation in microevolution, with the additional requirement that changes resulting from the operation of a particular mechanism of evolution could be rationalized in terms of the underlying transmission genetics (Integrated Knowledge Claim 1). The problem was constrained to two different mechanisms of evolution operating over time on the tails and collars of a single known population of lizards. The investigation activity was constrained to a limited field investigation with a teacher present to provide help and guidance. Both learning scenarios required many segments to accommodate the conduct of even a limited scientific investigation. Within each segment there were multiple tasks designed to target directly the disciplinary knowledge or inquiry skills involved so that the required evidence could be generated to support appropriate task-based feedback. Each segment begins by giving the student information to carry the problem forward from that point, regardless of their performance on past segments. Students could be moved from one segment to the next in sequence, or they would be able to choose segments at will.

Out of these learning scenarios grew the design for two culminating testlets. One culminating testlet was created as a compressed version of the experimental investigation in agouti mouse transmission genetics in the first learning scenario (Working Knowledge Claim 9). Compressed means that, whereas the learning scenario proceeeded very carefully and systematically over many segments to lead the student sequentially through each step of the investigation, the culminating testlet used the same organism in the same kind of investigation but required that the student choose and sequence the steps in the investigation.

The second culminating testlet was designed around an Integrated Knowledge claim by elaborating on use of knowledge representations presented in a single segment of this same learning scenario. Specifically, students would be required to produce evidence that they could understand the sexual life cycle model (which describes processes in an organism) in terms of models at lower levels of organization (e.g., ones relevant to cellular processes). Students would be moved from one testlet to the next.

When we finished our work, we had in fact created collections of task paradigms expressing purposes and relationships that could be bootstrapped and reused as the basis of design for other assessment tasks in either this or some other domain. Our domain experts agreed that the task design was consistent with their understanding of the standards. Beyond that, they thought that these kinds of tasks addressed parts of the standards most difficult to teach and, therefore, most underemphasized in the classroom.

Now we need to return to evidence paradigms.

2.5.5 Completing the Evidence Paradigms

Once we had completed our prototype task paradigms to the level described above, it was possible to continue with the completion of the bridges between proficiency and task paradigms – the evidence paradigms. Remember that these face in two directions and need information from both as well as supply

information to both. The design objects we use to make these connections are rules of evidence; that is, evaluation rules and interpretation rules. Evaluation rules (e.g., rubrics, response scoring algorithms) are designed to identify salient features of student work and evaluate them to produce values for our observables. Therefore, evaluation rules of evidence specify the connection between evidence and tasks. Interpretation rules specify how observables relate to aspects of proficiency – that is, how the evidence represented by an observable supports one or more claims related to one or more aspects of proficiency. Therefore, interpretation rules of evidence specify the connection between evidence and proficiency to be measured.

2.6 Looking Ahead to Student and Evidence Models and to Tasks

The real work of assessment design in terms of specifying **meaning** happens in the development of the paradigms we have described above. That is, meaning for the inferences we want to make, the evidence required to support those inferences, and the kinds of tasks that provide appropriate situations for collecting such evidence. It is at this level that the substantive argument of an assessment is laid out. Subsequent phases of design entail the refinement of these paradigms for the purposes of 1) satisfying particular product implementation and delivery constraints, and 2) quantifying our measurement models to reflect how we will update our beliefs about student proficiency given the quality of specific evidence gathered in specific implementations of specific situations. In sum, once we have addressed the assessment argument's meaning through the development of paradigms in domain modeling, we must use those paradigms as the semantic basis for developing models that represent the **machinery** of the assessment. At this point, it suffices to point out the basic relationships between our paradigms and the formal Biomass models they gave rise to.

2.6.1 The Student Model

Figure 17 depicts the Biomass Student Model, a part of the machinery for accumulating evidence to support claims about students. This Student Model is a fragment of a Bayes net – disccused in detail in the next chapter – which is the underlying psychometric model. Each node represents some aspect of knowledge or skill about which we wanted to accumulate evidence. They are derived directly from our conceptual representation of knowledge in the domain. There are sets of nodes that concern disciplinary, or declarative, knowledge; working knowledge, which is putting inquiry skills (such as reasoning through models) to use in the context of some disciplinary knowledge; and integrated knowledge, or reasoning through connections between models at different levels or for different but related phenomena. Each of the claims we articulated as important to make about students in either the interim or culminating assessment is related to one or more of these variables. That is, each claim is **about** a particular kind of knowledge or skill, or about some combination of them.

Fully specifying the Student Model in this phase of design includes definiting the relationships between the claims of interest and Student Model variables. It is the nature and grain-size of the discretely supported claims we want to make that determines which variables must be included in the Student Model. This Student Model can be used to accumulate evidence that backs the more detailed claims required by the interim assessment, as well as the more general claims for the culminating assessment. It is a re-usable design object. It has been built to store evidence about students that support the claims and purposes of Biomass – but it can be adapted for assessing inquiry in other domains and products, to the extent that the same general structure is a useful way of organizing what we want to say about students.

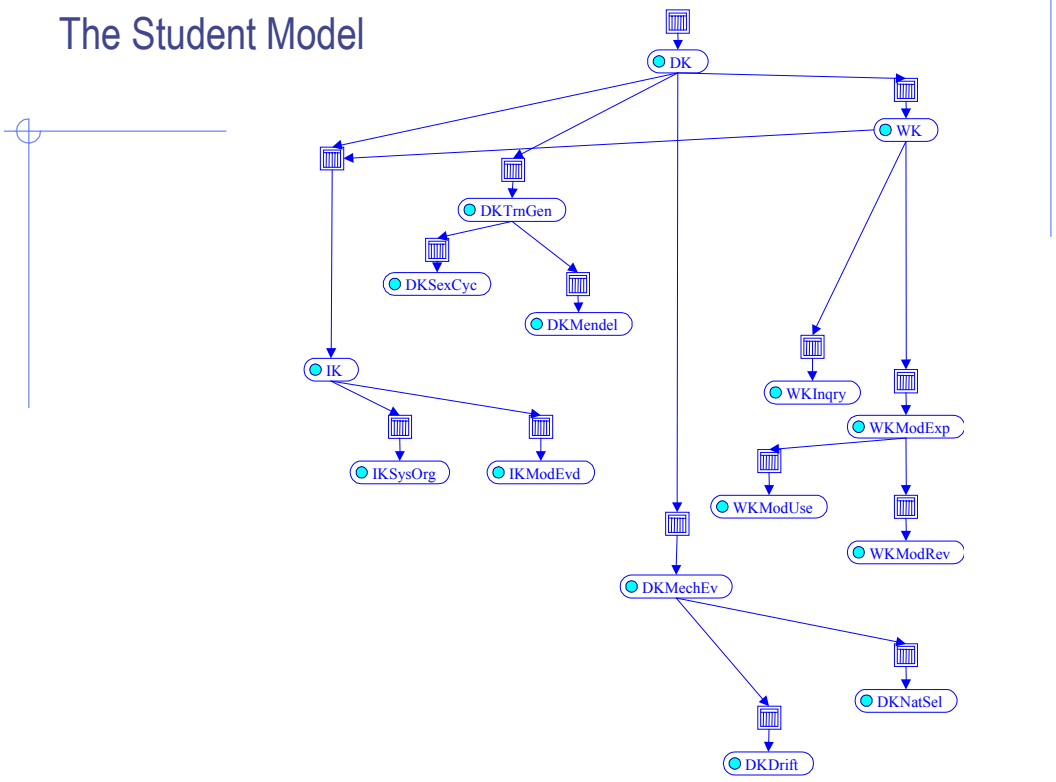

Figure 17. The Biomass Student Model

2.6.2 The Evidence Models

Figure 18 depicts the statistical fragments in some of the Biomass evidence models. They are used to take evidence in the form of the observables previously discussed, and update what we know about student model variables. These evidence models are also re-usable objects. They can be used to update the Student Model in recurring, structurally-similar, situations, in which certain kinds of evidence is obtained to tell us about certain aspects of knowledge and skill. The same fragment might be used with many complex tasks that are all built around the same task model. In Biomass we used several of these evidence models more than once – same structure, but with different individualized parameters. As with the Student Model, these mathematical models – the machinery – must satisfy the requirements not only of the substantive assessment argument, but also various operational constraints (e.g., re-use, performance).

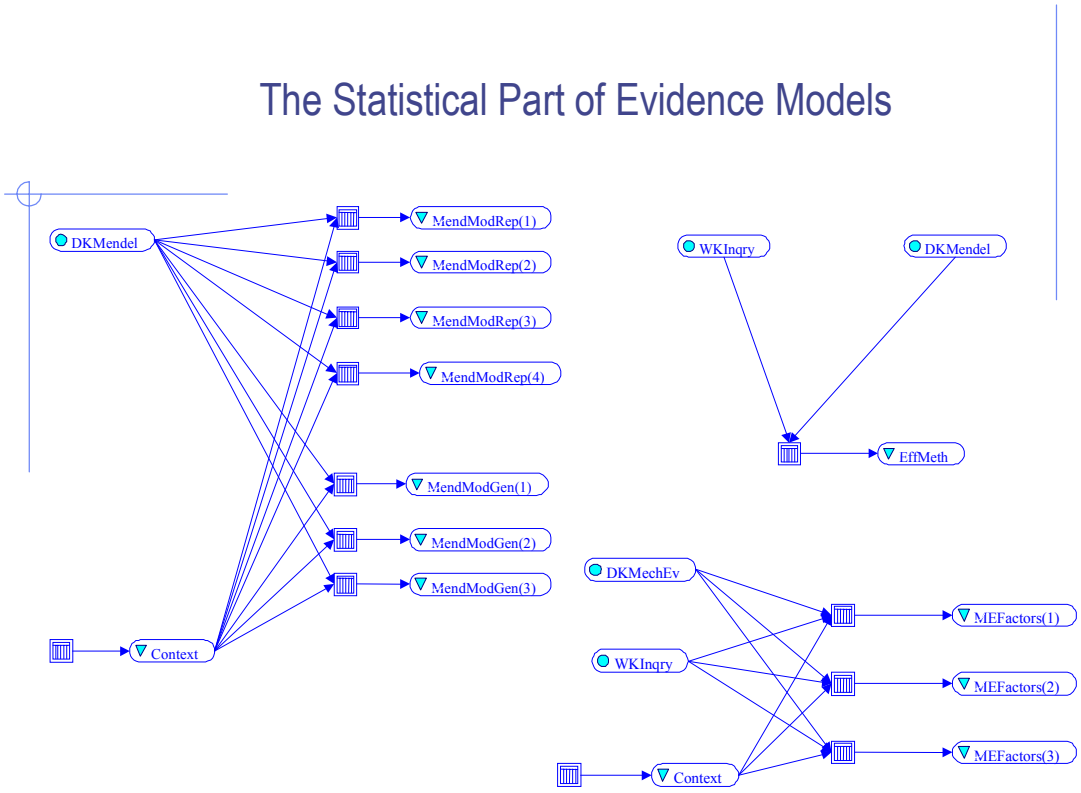

Figure 18. The Statistical Part of Selected Evidence Models

2.6.3 Tasks

Figure 19 is an example of a specific task generated from a task model. The task shown is an implemented instance derived from a more general schema described by a task model. Such a task model would include variables specifying the number and nature of the representational forms (e.g., four kinds of dominance relationships) to be made available to the student for expressing whatever the hypothesis happens to be, as well as variables specifying the elements of the hypothesis itself. A task model (as opposed to a task paradigm) is also specialized to describe operational task presentation requirements (in this case a Web-based drag/drop task). That is, a task *paradigm* provides at a more narrative level the structure of a family of tasks, while a task *model* provide specifications for implementing such tasks and ensuring their operational elements will be compatible with those of evidence models. The use of task models, with their variables and specifications for task materials, guides task authoring that is intended to achieve a particular evidentiary focus and level of difficulty, while reducing construct-irrelevant factors to the greatest extent possible.

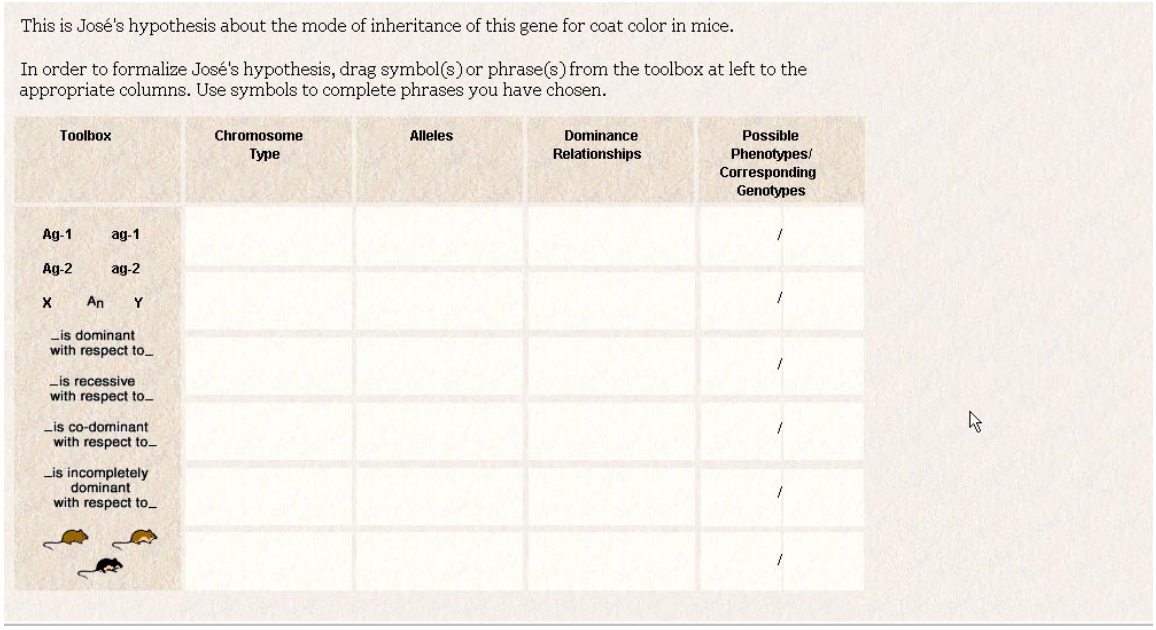

Figure 19. The "Mode of Inheritance Table" Representational Task, before Responses

2.7 Implications for the Assessment Delivery Processes

As part of the Biomass project's capability demonstration, we intended to implement a first version of the Four Process Assessment Delivery System pictured below in Figure 20. (In this figure the major assessment delivery processes appear at the corners of the diagram. The cloud-like objects connecting them are data.) The implications of evidence-centered design for assessment delivery system architecture in general are quite powerful. (A more detailed accounting is provided in *A Four Process Architecture for Assessment Delivery, with Connections to Assessment Design*, Almond, Steinberg, & Mislevy, in press).

The generic ECD models themselves and the manner in which they can be combined and recombined to meet new purposes and conditions provide a fresh perspective on assessment delivery processes. In the traditional delivery paradigm related to traditional high-stakes testing with multiple choice items, delivery processes are typically 'bundled' to accommodate item types. A single process does all the work of presenting multiple choice response items rendered in one or more particular formats (e.g., radio button selection of a single response or multiple responses), scoring the response as right or wrong, then adding one to number right. This means that if any part of the 'bundle' changes (e.g., rendering, response scoring or summary scoring), the whole process has to be replaced.

What is accomplished by using a more flexible architecture is that unbundling the processes enables a 'plug and play' delivery strategy. This means, for example, that the same material can be presented in two different assessments (applications) using the same Presentation Process but use completely different Evidence Identification Processes for response scoring (one might be right/wrong and the other diagnostic). These two applications may then converge again in using a number right Evidence Accumulation Process, or diverge further with one using IRT and the other a Bayes Net inference engine for Evidence Accumulation. Any particular instance of a delivery system is configured to use the correct subset of all processes available. The data maintained in the Task/Evidence Composite Library (discussed in more detail below) support the crucial links between processes. The use of ECD in combination with the assessment delivery architecture illustrated here is very compelling because together they represent a coherent and universally applicable path leading from design to delivery. Let us look at each of these processes in turn and consider the implications of the Biomass design so far.

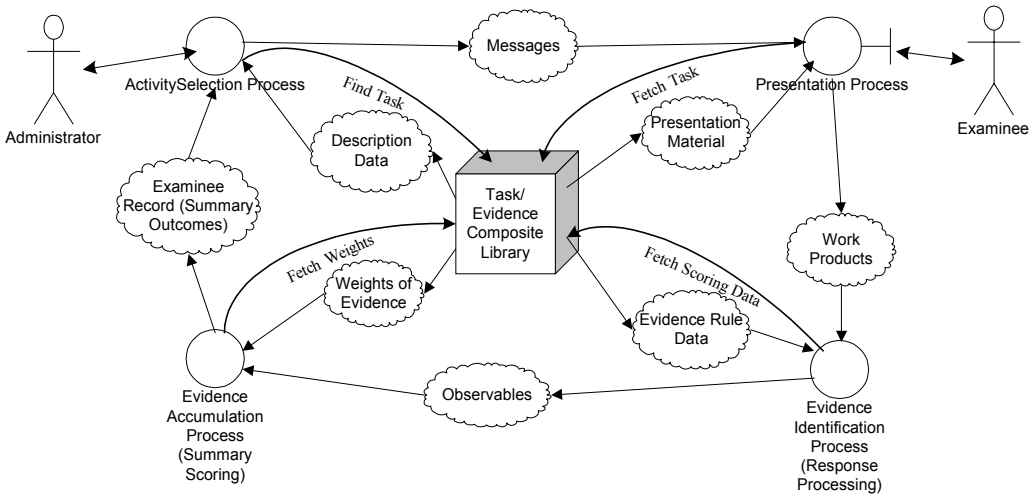

Figure 20: The Four Process Assessment Delivery System

2.7.1 Presentation Process

The PRESENTATION PROCESS is responsible for presenting the task to the examinee. It gets information about the task from the task library as necessary. When the examinee performs the task, the PRESENTATION PROCESS will capture one or more Work Products (i.e., results) from the examinee. It sends them to the EVIDENCE IDENTIFICATION PROCESS to be evaluated. The knowledge representations we present are the Presentation Material; the knowledge representations we capture are the Work Products. An important feature of this delivery architecture is that it separates task content from any particular rendering of that content. This means that task material may be designed to be presented by different Presentation Processes. A good example of this are the many presentation incarnations possible for an item producing a work product containing a multiple choice response. The same content may be rendered in one kind of Presentation Process where the student uses high-lighting to make a response and in another kind of Presentation Process where the student clicks on a radio button to make a response. In either case the content of the material presented and the work product captured are the same.

An important requirement for Biomass was that it be delivered over the Web, with corresponding requirements for the Presentation Process. The Biomass tasks could, however, be adapted to a different Presentation Process, such as paper and pencil.

2.7.2 Evidence Identification (aka Response Scoring)

The EVIDENCE IDENTIFICATION PROCESS, or response scoring, carries out the first step in the scoring process: It identifies the essential features of the response that provide evidence about the examinee's current knowledge, skills, and abilities. These are recorded as a series of Observations (that is, values of Observable Variables). These can be passed to the next process, EVIDENCE ACCUMULATION or summary scoring (see below), used to trigger feedback to the examinee immediately, or both. Biomass was intended to support learning as well as high stakes selection. To that end we had designed both an interim and a culminating assessment. These different versions each had its own collection of tasks, as we have described above; any given task could produce a complex work product requiring specific evaluation rules. For the kind of Web-based delivery envisioned, all of these rules would have to be automated within at least two Evidence Identification Processes – one that would be used with the interim assessment and one that would be used with the culminating assessment.

However, we had decided that students should be able to use interim tasks in two ways: as a means of practice with task-based feedback, or as a means of practice without task-based feedback. This had

additional consequences for the Evidence Identification Process because it meant that any work product produced by an interim task could be evaluated in two different ways: degree of correctness and diagnostically. Which way student work was evaluated depended on whether or not the student asked for feedback during a segment of the learning scenario. This meant we needed three Evidence Identification Processes: one for the culminating mode and two for the interim mode. The observables produced by the diagnostic process would not be passed along to the Evidence Accumulation Process, but would only be used to trigger feedback to the student specifically about her performance on the task at hand.

2.7.3 Evidence Accumulation (aka Summary Scoring)

The EVIDENCE ACCUMULATION PROCESS, or summary scoring, performs the second stage in the scoring process: synthesizing evidence across multiple tasks in terms of their implications for the examinee's student-model variables. That is, it updates our beliefs about the examinee's knowledge, skills, and abilities based on the evidence from each successive task.

In Biomass, the requirement for providing summary scores on multiple aspects of proficiency meant that the Evidence Accumulation Process would be implemented as a Bayes Net. As observables from each task are determined by the Evidence Identification Process, they are passed on to the Evidence Accumulation Process. When the particular form this latter process takes is a Bayes net, the evidence-model Bayes net fragments (as depicted in Figure 18) are "docked" one at a time with the student-model Bayes net fragment (as depicted in Figure 17). Evidence about the student model variables is then used to update the probability distribution in the Scoring Record that summarizes belief about the values of the student-model variables (Almond & Mislevy, 1999; Mislevy & Gitomer, 1996; Mislevy et al., in press). Score reports would include information that related student model variable statistics with the claims used to interpret them. The function used to generate a reporting statistic from the Scoring Model (an instance of the student model maintained and updated for a given student, as her responses arrive) would be the same for both culminating and interim assessments; however, the student model variables sampled for the culminating assessment would only be a subset of all those in the model.

In Biomass there are a number of different observables with different combinations of student model variable parents coming from different tasks. As mentioned above, we increased re-use and reduced implementation burden by identifying a number of structurally similar relationships among student-model and observable variables. We were able to build Bayes net fragments around these patterns, and employ many of them repeatedly. The statistical portion of an evidence model includes both the *structure* of the relationship between each observable and its student model parent(s) and the *strength* of that relationship. Therefore, if we could identify crucial patterns of relationships and reuse them not only would our implementation time be reduced, but we also would have created a collection of evidence models that could be reused in completely different assessment applications. The strength of the relationships would be estimable from combinations of task features, expert opinion, and emprical pretest data (as discussed in Mislevy, Almond, Yan, Steinberg, 1999).

2.7.4 Activity Selection

The ACTIVITY SELECTION PROCESS is responsible for selecting a task from the Task Library. These could be tasks with a focus on assessment or on instruction, or they could be activities related to test administration. In an adaptive system, the ACTIVITY SELECTION PROCESS may consult what is currently known about the examinee (contained in the Examinee Record) to decide when to stop, what kind of task to present next, or to present an instructional task as opposed to an assessment task. Because Biomass was a prototype and not an operational product, we wanted maximum flexibility in how tasks were selected for presentation. Therefore, we decided that the Activity Selection Process would be linear (get next) for the culminating assessment and student-driven for the interim assessment. In the interim mode students would be given the option of going on to the next segment in the learning scenario or to any other one they wanted.

2.7.5 Task/Evidence Composite Library

Sitting in the middle of Figure 20 is the Task/Evidence Composite Library. This library (which can be imlemented in any number of ways) makes available to the four processes information that they need to carry out their functions. As indicated in Figure 21, it contains 1) the implemented version of materials for any given task, linked with 2) the implemented evidence model(s) required to handle the evidence they produce, plus 3) fixed categories of data designed to support the various processes in assessment delivery.

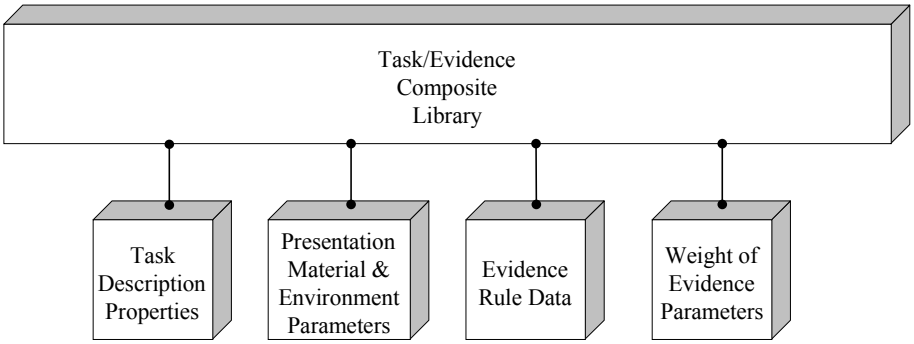

Figure 21. The Task/Evidence Composite Library

2.7.6 Controlling the flow of information among the processes

One of the critical differences between Four Process Delivery architecture and more traditional delivery systems is the way in which the flow of information through the system is managed. In traditional delivery, this flow is usually 'hard-wired'; that is, because of the assumptions inherent in assessment for selection, delivery processes are strictly defined to receive fixed information from and send fixed information to certain other processes in a fixed sequence. When you change the purpose of an assessment -- its context or conditions for use-- all this needs to be able to change as well. The two modes of Biomass prototype illustrate the differing logic requirements for the sequence of interactions among the delivery processes.

The purpose of the Culminating Assessment is to determine a student's level of proficiency at the end of a course, providing overall results and summary feedback at the end of the testing session. The ACTIVITY SELECTION PROCESS tells the PRESENTATION PROCESS to start each successive task after capturing the work products of the previous one. All of the work products can be sent to EVIDENCE IDENTIFICATION at once for task-level scoring. EVIDENCE IDENTIFICATION can use the resulting observable variables in two ways. Some observable variables can be used to generate task-level feedback, which may be presented to the examinee immediately or at the end of the assessment, as appropriate to the assessment's purpose. And some observable variables can be passed on to EVIDENCE ACCUMULATION to update beliefs about student-model variables, and subsequently ground higher-level feedback, instructional decisions, or score reports. Note that all of the task- and summary-scoring scoring can be accomplished by EVIDENCE IDENTIFICATION and EVIDENCE ACCUMULATION at a distant time or place from the actual testing session.

The purposes of the Interim Assessment, on the other hand, are to provide practice and support learning in preparation for the Culminating Assessment. These purposes are served by immediate feedback, cumulative scoring, and opportunity to repeat a task. Now the delivery process cycles around the entire outer ring of the diagram for each task. The ACTIVITY SELECTION PROCESS tells the PRESENTATION PROCESS to start a new task. The student produces work products, which are sent to EVIDENCE IDENTIFICATION for task-level scoring and task-level feedback. Viewing the feedback, the student either decides to repeat the task, so ACTIVITY SELECTION tells the PRESENTATION PROCESS to administer the same task again, or decides to move on. In this case, EVIDENCE ACCUMULATION updates beliefs about the student. This information is available to trigger instruction, provide interim feedback, select further activities, or display provisional or final scores, all as appropriate to the purpose of the assessment

The Four Process Delivery Architecture achieves the desired flexibility not only by supporting 'plug and play' for Processes, but also by controlling the flow of information among them with messaging that's customizable for any given assessment application.

Figure 21represents a system's view of all these requirements.

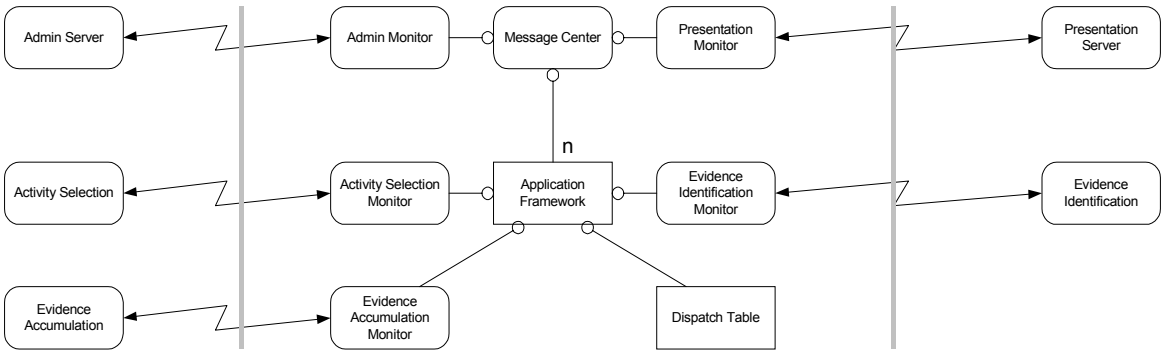

Figure 21. A System's View of Four Process Assessment Delivery

CONCLUSION

Biomass is a prototype assessment that illustrates several innovative and ambitious features, including the following:

- Designing assessments in terms of re-usable schemas, objects, and processes.
- Developing assessments to assess standards, in such a way as to both give them concrete meaning and address higher-level forms of knowledge--in this case, inquiry in science, with content from transmission genetics and microevolution.
- Using dynamically-assembled Bayesian inference networks to manage the accumulation of evidence in a multivariate model, from multivariate and sometimes conditionally dependent observations (as in Bradlow, Wainer, & Wang, 1999).
- Using innovative technologies in a way that flows naturally from the evidentiary requirements of an assessment's purpose--in this case, using web delivery and automated scoring in the service of learning tasks and culminating tests.

This presentation has shown how the models and approaches of evidence-centered design can be used to organize the design and implementation of such an assessment, and do so in a way that lends itself to the re-use of the materials and processes.

REFERENCES

American Association for the Advancement of Science (1993). *Benchmarks for Scientific Literacy*. Oxford, UK: Oxford University Press.

American Federation of Teachers. (1995). Making standards matter: A fifty-state progress report on efforts to raise academic standards. Washington, DC: Author.

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-237.

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (in press). A four-process architecture for assessment delivery, with connections to assessment design. *Journal of Technology, Learning, and Assessment*.

American Association for the Advancement of Science (AAAS, 1994). Benchmarks for Scientific Literacy.

BSCS (Biological Sciences Curriculum Study) (1993). *Developing Biological Literacy: A Guide to Developing Secondary and Post-secondary Biology Curricula.* Colorado Springs, CO: Author.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). "A Bayesian random effects model for testlets." *Psychometrika, 64*, 153-168.

Descotte, Y., & Latombe, J-C. (1985). Making compromises among antagonist constraints in a planner. *Artificial Intelligence, 27,* 183-217.

Katz, I. R. (1994). Coping with the complexity of design: Avoiding conflicts and prioritizing constraints. In A. Ram, N. Nersessian, & M. Recker (Eds.), Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Kindfield, A.C.H. (1999). Generating and using diagrams to learn and reason about biological processes. *Journal of the Structure of Learning and Intelligent Systems, 14(2),* 81-124.

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

Mislevy, R.J, Almond, R.G., Yan, D. and Steinberg, L.S. (1999) "Bayes Nets in Educational Assessment: Where the numbers come from." In Laskey, K.B. and Prade, H. (eds.), *Uncertainty in Artificial Intelligence '99* (437-446). Morgan-Kaufmann.

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). On the roles of task model variables in assessment design. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). Hillsdale, NJ: Erlbaum.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (in press). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives.*

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (in press). Making sense of data from complex assessments. *Applied Measurement in Education.*

National Council of Teachers of Mathematics. (1989). *Curriculum and performance statndards for school mathematics.* Reston, VA: Author.

National Research Council (1996). *National Science Education Standards.* Washington: National Academy Press.

New Standards Project. (1997). *Performance standards for high school English Language Arts, Mathematics, Science, and Applied Learning*. Oxford, UK: Oxford University Press.

Stewart, J. & Hafner, R. (1991). Extending the conception of "problem' in problem-solving research. *Science Education, 75(1),* 105-120.