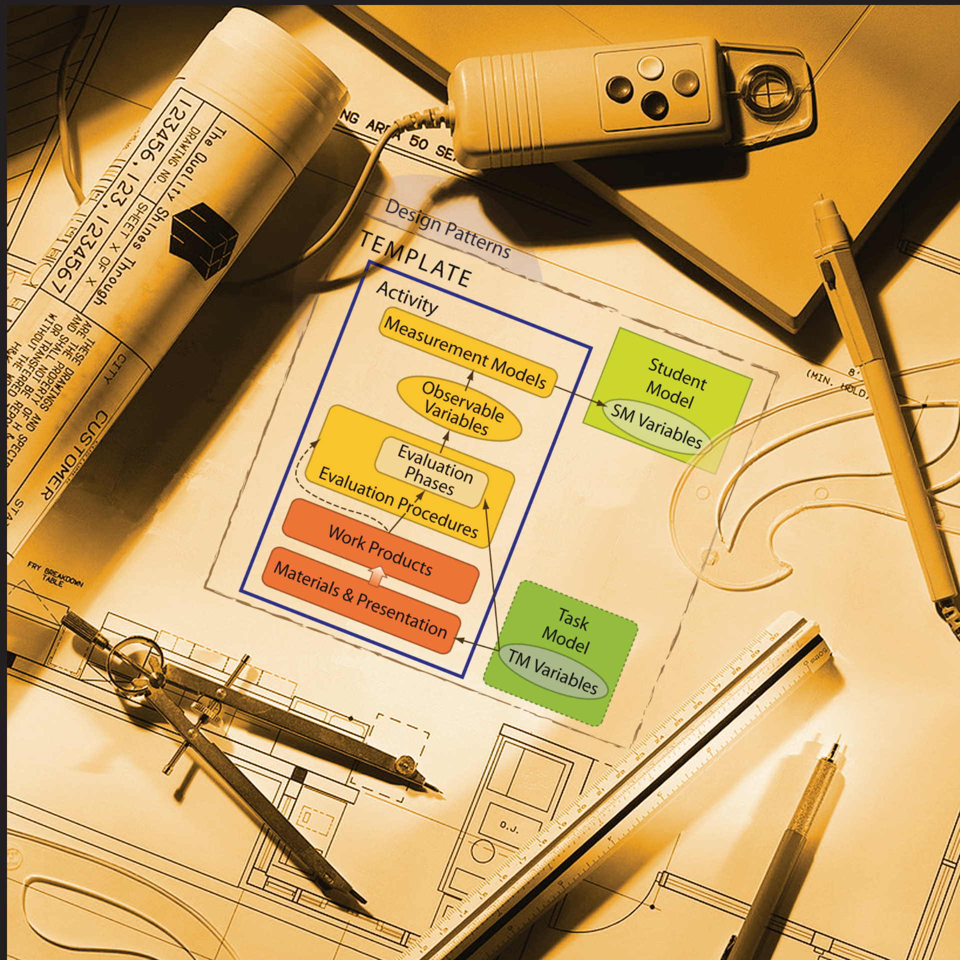


Large-Scale Assessment Technical Report 6| May 2009

Design Patterns for Assessing Model-Based Reasoning

Project: Application of Evidence-Centered Design to State Large-Scale Science Assessment

Robert J. Mislevy, University of Maryland
Michelle M. Riconscente, University of Southern California
Daisy Wise Rutstein, University of Maryland

Report Series Published by SRI International

SRI International
Center for Technology in Learning
333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
http://ecd.sri.com

Technical Report Series Editors
Alexis Mitman Colker, Ph.D., *Project Consultant*
Geneva D. Haertel, Ph.D., *Principal Investigator*
Robert Mislevy, Ph.D., *Co-Principal Investigator*
Ron Fried, *Documentation Designer*

Application of Evidence-Centered design to state large-scale science assessment

Technical Report 6

Design Patterns for Assessing Model-Based Reasoning

May 2009

Prepared by:
Robert J. Mislevy, University of Maryland
Michelle M. Riconscente, University of Southern California
Daisy Wise Rutstein, University of Maryland

Acknowledgments
This material is based on work supported by the National Science Foundation under grant DRL-0733172 (An Application of Evidence-Centered Design to State Large-Scale Science Assessment).

Contents

FIGURES

List Of Tables

List Of Boxes

ABSTRACT

Understanding, exploring, and interacting with the world through models characterizes science in all its branches and at all levels of education. Model-based reasoning is central to science education and thus science assessment. Building on research in assessment, science education, and learning sciences, this report provides a set of *design patterns* to help assessment designers, researchers, and teachers create tasks for assessing aspects of model-based reasoning: Model Formation, Model Use, Model Elaboration, Model Articulation, Model Evaluation, Model Revision, and Model-Based Inquiry. Each *design pattern* lays out considerations concerning targeted knowledge and ways of capturing and evaluating students' work. The ideas are illustrated with examples from existing assessments and the research literature.

1.0 Introduction

Models are fundamental to science. The centrality of Newton's laws, the double helix model of DNA, and the Lotka-Volterra model of predator-prey interaction are cases in point (Frigg & Hartmann, 2006). But it is not simply the content of models that matters. Scientists build, test, compare, and revise models. They use models to organize experience, guide inquiry, communicate with one another, and solve practical problems (Lehrer & Schauble, 2006). The National Science Education Standards (NSES; National Research Council, 1996) highlights "Evidence, Models, and Explanation" as a unifying theme for science education, spanning grade levels and science domains. To guide and evaluate students' progress, it is therefore important to be able to assess their proficiencies in reasoning with and about models.

While it is fairly straightforward to assess students' familiarity with terminology and calculation, assessing model-based reasoning is more challenging (National Research Council, 2001). How can we devise occasions and settings for students to display their capabilities to build, critique, revise, and use models in order to understand, explain, predict, and produce effects in the natural world? How might we evaluate the cycles of observing, hypothesizing, and reformulating that characterize inquiry using models? How can we leverage technology to better assess model-based reasoning? When we are constrained by time and costs, can we craft simpler tasks that nevertheless provide some evidence about key aspects of model-based reasoning? Are there principles and approaches to help us assess model-based reasoning across the diversity of models used in different branches of science and across levels of education from the primary grades to post-secondary study?

This paper provides support for designing tasks that assess model-based reasoning in the form of a suite of *design patterns. Design patterns* are used in architecture and software engineering to characterize recurring problems and approaches for solving them such as Workplace Enclosure for house plans (Alexander, Ishikawa, & Silverstein, 1977) and Interpreter for object-oriented programming (Gamma, Helm, Johnson, & Vlissides, 1994). *Design patterns* for assessment likewise help domain experts and assessment specialists "fill in the slots" of an assessment argument built around recurring themes in learning (Mislevy et al., 2003). The particular form of *design patterns* presented here were developed in the Principled Assessment Design for Inquiry (PADI) project (Mislevy et al., 2003).

The following overview of model-based reasoning draws on Stewart and Hafner's (1994) and Gobert and Buckley's (2000) analyses of model-based reasoning in science. These authors highlight the importance of interactivity and iteration in the ways scientists use models—continually constructing and reconstructing correspondences between general structures and unique real-world situations. Table 1 lists the *design patterns* that are addressed in the paper. They can be used one at a time to develop tasks that target particular aspects of model-based reasoning or used jointly to develop more complex multi-stage or iterative investigations. A summary of the "evidence-centered" approach to assessment under which *design patterns* are conceived is then presented, along with a description of the attributes of a PADI *design pattern*.[1]

The remainder of the report presents the *design patterns*. One iterative investigation task, based on a genetics investigation in a curriculum devised by Stewart and his colleagues (Johnson & Stewart, 2002) appears at several points in the presentation, showing how the *design patterns* can contribute in combinations to ground complex tasks. The introduction to this genetics investigation is given in Box 1. More focused tasks specific to particular aspects of model-based reasoning appear in the discussions of each of the *design patterns*. A variety of models, content domains, task types, and educational levels are used in the examples to indicate the breadth of applicability of *design patterns* as a starting point for task design.

The conclusion summarizes the rationale for the use of *design patterns* to support developing tasks to assess model-based reasoning in science. We note their relevance to standards-based assessment, instruction, and large-scale accountability testing.

Table 1: Aspects of Model-Based Reasoning in Science

Aspect	Definition
Model Formation	Establishing a correspondence between some real-world phenomenon and a model, or abstracted structure, in terms of entities, relationships, processes, behaviors, etc. Includes determination of the scope and grain-size to model, which aspects of the situation(s) to address and which to leave out.
Model Use	Reasoning through the structure of a model to make explanations, predictions, conjectures, etc.
Model Elaboration	Combining, extending, and adding detail to a model. Establishing correspondences across overlapping models into larger assemblages. Fleshing out more general models with more detailed models.
Model Articulation	Connecting meaning of physical or abstract systems across multiple representations. Representations may take qualitative or quantitative forms. Notably relevant in models with quantitative and symbolic components, such as the conceptual and mathematical aspects of physics models.
Model Evaluation	Assessing the correspondence between the model components and their real-world counterparts with emphasis on anomalies and important features not accounted for in the model.
Model Revision	Modifying or elaborating a model for a phenomenon in order to establish a better correspondence. Often initiated by model evaluation procedures.
Model-Based Inquiry	Working interactively between phenomena and models, using all aspects of the above. Emphasis on monitoring and taking actions with regard to model-based inferences vis-à-vis real-world feedback.

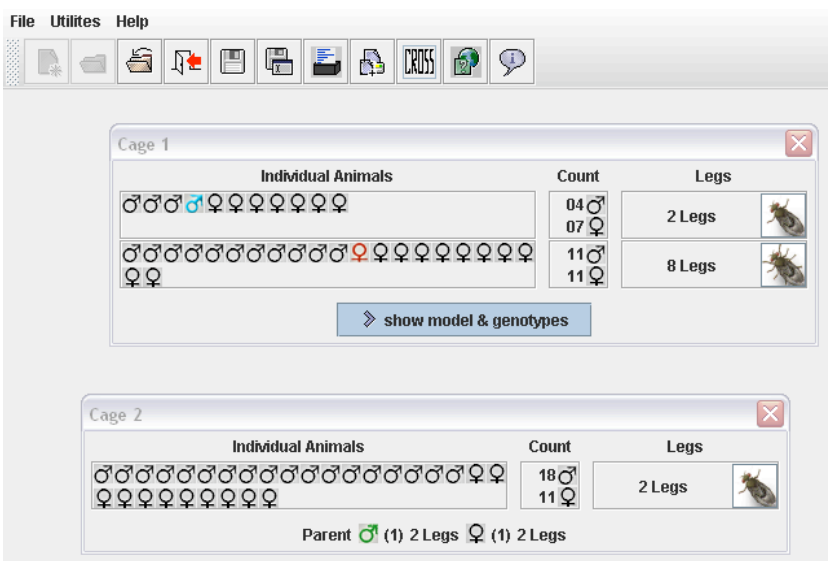
[1] The reader interested in fuller discussions of evidence-centered design is referred to Mislevy, Steinberg, and Almond (2003), Mislevy and Riconscente (2006) and the series of technical reports from the Principled Assessment Design for Inquiry (PADI) project (http//padi.sri.com/publications.html).

Box 1: Model-Based Reasoning Tasks in Genetics: Introduction

Stewart and Hafner developed a course containing laboratories for the study of baseline genetics models. These laboratories included the use of Jungck and Calley's *Genetics Construction Kit* (GCK; 1985), a software simulation program that includes the ability to construct customized problems to study different genetics phenomena (Stewart, Hafner, Johnson, & Finkel, 1992).

At the start of the course, students learned about the development of models and read an abridged version of Mendel's paper (Johnson & Stewart, 2002). They were then visited by a graduate student dressed as Mendel who taught them about the simple dominance model. Students were subsequently tasked with using the GCK to determine whether the simple dominance model could be applied to test crossing organisms. Using a box containing several specimens with a given trait, students performed a cross and examined the resulting output to identify which crosses produced which traits. They then responded to questions regarding their findings, such as which trait appeared to be the most dominant. The figure below, which represents an early stage in problem-solving, is taken from the Virtual Genetics Kit, software based on the GCK (http://intro.bio.umb.edu/VGL/index.htm).

The initial task represented a case of model use, in which students applied a known model to a given set of data. The next task presented data that did not strictly adhere to the simple dominance model in order to advance students' understanding of model-based reasoning. Students were prompted to evaluate the fit of the data to the known model and, upon discovering that the known model was inadequate, to revise the existing model to account for the observed deviations. Working in groups, students conducted their analyses and developed revised models they tested using crosses provided by the other groups. Each group then presented their solution. This process was repeated for different models, such as the codominance and the multiple allele models. Data were collected from each round and used to assess each student's proficiency in model revision. These data included recordings of the research group interactions both internally and with the instructor; lab books; interactions with the software such as the sequence of actions performed; and a written description of the group's final model. Later sections of this paper provide details on additional examples and associated assessments of student ability.



Screen shot from the Virtual Genetics Kit

2.0 Model-Based Reasoning

Research into the processes of learning yields insights into characteristics of effective instruction, such as the importance of active engagement in the learning process and the integrated development of declarative, procedural, conceptual, and social aspects of knowledge (National Research Council, 2000). In the case of science, researchers stress the importance of involving students in the inquiry processes that scientists apply in their own work, rather than teaching only the conclusions scientists have reached (Stewart, Hafner, Johnson, & Finkel, 1992). Broadly speaking, inquiry can be thought of as the process by which investigators formulate and investigate questions about the natural world in order to formulate answers, explanations, predictions, or theories (NSES). The development of inquiry skills involves more than recalling facts and terminology; it means being able to reason through fundamental concepts and relationships to understand and interact with particular real-world situations—in short, model-based reasoning.

2.1 Scientific Models

Researchers such as Stewart and Gobert unify thinking about science content, the process of inquiry, and teaching and learning in science in terms of model-based reasoning. A model is a simplified representation focused on certain aspects of a system (Ingham & Gilbert, 1991 as cited in Gobert & Buckley, 2000). The entities, relationships, and processes of a model constitute its fundamental structure. They provide a framework for reasoning across any number of unique real-world situations, in each case abstracting salient aspects of those situations and going beyond them in terms of mechanisms, causal relationships, or implications at different scales or time points that are not apparent on the surface.

This presentation focuses on the explicit models that are the object of science instruction (which we will see also entails aspects of reasoning through mental models; Johnson-Laird, 1983). A scientific model can be viewed as a community resource—a special and particularly technical case of what cognitive anthropologists call cultural models (Strauss & Quinn, 1998). The concepts, relationships, and processes constitute a system that has an existence beyond the mind of any one individual. It is manifest in books and tools, in ways of seeing the world, and in patterns of interacting with the world and other people. A web of interrelated ideas and activities spans individuals, is contributed to by many, is used by many more, and is enriched with use. Individuals vary with respect to their depth, breadth, areas, and capabilities with a given model, in view of their unique histories of experience, and no one individual may command the totality of concepts, tools, and applications for a complex model. The goal of science education, according to NSES, is to bring students into the community—to acquaint them with key concepts and relationships of important models, to be sure, but further to empower them to interact with the ideas and with people in practically useful ways that are mediated by scientific models.

A broad conception of models serves present purposes, in order to highlight similarities in kinds of thinking that apply across a broad range of models (see Frigg & Hartmann, 2006, for an overview of models in science). We want to ground *design patterns* on overarching similarities in order to guide task design across a broad range of assessment situations. More specialized *design patterns* can be constructed for more specialized classes of models and representations and would provide more focused support for particular areas of science or kinds of tasks. For our purposes, models can be as simple as the change, combine, and compare schemas in elementary arithmetic (Riley, Greeno, & Heller, 1983), or as complex as quantum mechanics, with its multiple forms of representation, advanced mathematical formulations, and rich interconnections with other physical models. Models can contain or overlap with other models. Relationships among entities in models may be qualitative, relational, dynamic, or spatial (Gobert, 2000). Some models concern processes, such as the stages of cell division in meiosis. Some relationships in models are qualitative (if Gear A rotates clockwise, Gear B must rotate counterclockwise), and some extend to quantitative or symbol-system representations and associated operations (if Gear A has 75 teeth and Gear B has 25 teeth, Gear B will rotate three times as fast as Gear A).

Figure 1 suggests some central properties of a model (Greeno, 1983; Mislevy, 2009). The lower left plane shows phenomena in a particular real-world situation. A mapping is established between this situation and, in the center, patterns expressed in terms of the entities, relationships, and properties of the model, or the "semantic" layer of the model. Reasoning is carried out in these terms. This process constitutes a reconception of the situation which synthesizes particulars of the situation with the abstracted structure of the model, akin to what Fauconnier and Turner (2002) would call a "blended space" for reasoning. The processes and relationships of the model are used to make inferences about the real-world situation, such as explanations, predictions, or plans for action (Swoyer, 1991). Above the plane of entities and relationships in the models are symbol systems that further support reasoning in the model space, such as the matrix algebra and path diagram representations used in structural equation modeling. (Working with model representations is the focus of Model Articulation, discussed in Section 8).

Figure 1: Reconceiving a Real-World Situation through a Model

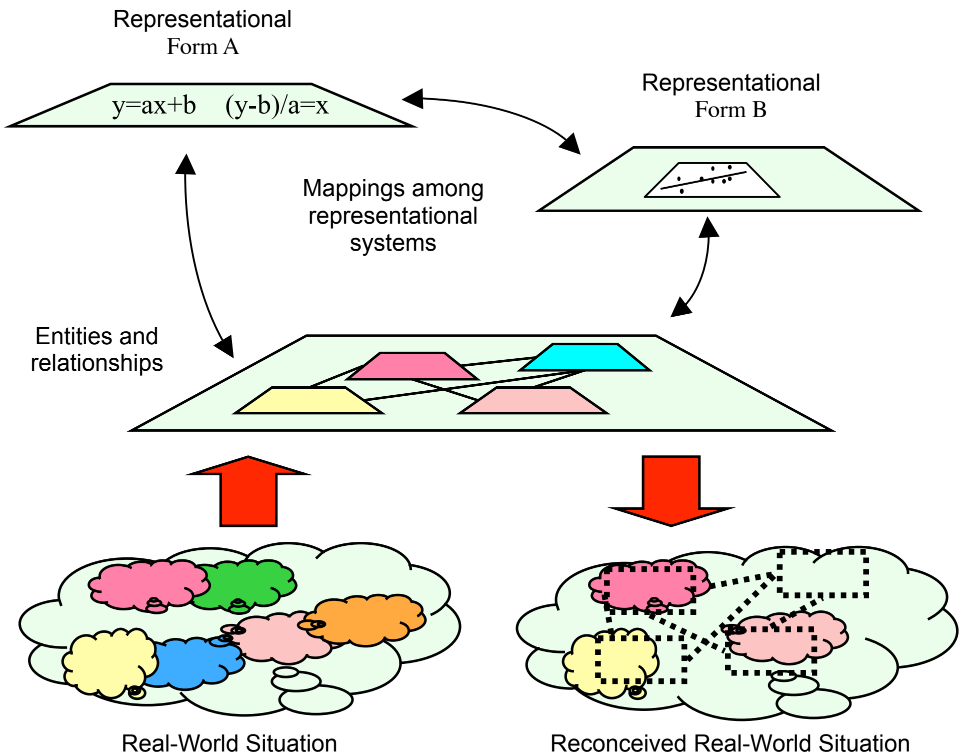


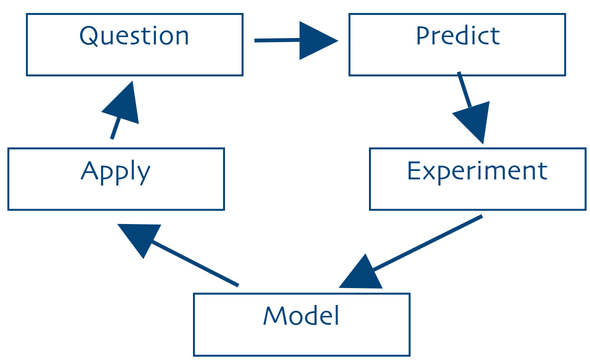
Figure 1 highlights some properties that are important for understanding how models are used. The real-world situation is depicted as nebulous, whereas the model is crisp and well defined. Not all aspects of the real-world situation have corresponding representations in the model. On the other hand, the model conveys ideas and relationships that the real-world situation does not. The modeled situation shows a less-than-perfect match to the model, but it provides a framework for reasoning that the situation itself does not. The question of a model's validity does not address a two-way relationship between a model and reality, but a four-way relationship among a model, reality, a user, and a purpose (Suárez, 2004).

The processes of proposing, instantiating, checking, and revising to find an apt model for a given purpose in a given situation thus characterizes model-based reasoning in practice. The reconceived understanding is typically provisional. Hypothesized missing elements can be used to evaluate the quality of the representation and prompt one to revise or abandon a particular model. The hypothesized relationships guide actions that change real-world situations and give rise to further cycles of inquiry, understanding, and action. Figure 1 does not convey the strategies, procedures, and rules of thumb that enable one to put a model to practical use. These are the "epistemic games" (Collins & Ferguson, 1993) that students must learn if they are to develop their capabilities for reasoning with models.

2.2 The Inquiry Cycle

Model-based reasoning lies at the heart of inquiry that uses models to formulate explanations from observations. In traditional science education, students are presented with models and asked to apply them to problems (Stewart & Hafner, 1991). A model-based reasoning approach extends the use of models in instruction to modifying existing models and developing new ones. Students may be presented—or propose on their own—a question that can be addressed by the concepts and principles in a scientific domain and then need to determine what observations might bear on its solution. They may be presented with—or gather themselves—data about the natural world and be prompted to build a model that can account for the patterns in the data. Once they have formulated a model, they may be asked to test the model by making predictions about further observations and to determine whether it holds up in light of new information or requires modifications—and if so, the cycle of model-building, model-checking, and model-revision will continue, each stage requiring its own particular kind of reasoning. White, Shimoda, and Frederiksen (1999) depict the inquiry cycle as shown in Figure 2. It provides a useful schema for discussing aspects of model-based reasoning as they are used in inquiry and as they can be addressed in assessment.

Figure 2: The Inquiry Cycle



Using models serves both pedagogical and functional goals in instruction. Typically, students are introduced first to simpler forms of models and inquiry (e.g., provided substantial scaffolding to guide their investigations) and then are exposed gradually to more complex models (as described in the genetics example in Box 1) and more independent situations for using them.

The multifaceted nature of model-based reasoning holds implications for both instruction and assessment. An instructor's decision to develop students' proficiency specific to one aspect will require assessment attuned to that aspect. The focus of instruction, and thus assessment, for a new model may initially be reasoning through that model with data for which it is known to be appropriate. Alternatively, an instructor may want to see students work through cycles of inquiry with a model with which students are already familiar, in order to focus on the monitoring and organizational capabilities required in iterative model-fitting.

Further, students do not develop competence across aspects of model-based reasoning at the same rate and depth. One student may formulate new models successfully but have trouble determining whether their models fit the data. The same student may be more facile with some aspects of inquiry in some content domains than others—and even for different investigations within the same domain. *Design patterns* for assessing model-based reasoning must therefore be able to support inference about targeted aspects of inquiry as well as their integrated orchestration.

Every application of model-based reasoning involves particular models at some level. Instructors and assessment designers need to think about the interplay between models and model-based reasoning and where they want to focus attention. For example, an exercise meant to highlight model-checking could use a model familiar to students. An exercise meant to expand students' capabilities with a new model could employ a model-checking technique students are familiar with from a previous lesson. The designer must consider the extent to which declarative knowledge of a model's structure and components—as opposed to reasoning with and through the model—are to be stressed. This determination depends not simply on what is in the task but also on the relation of that task to the experience of the examinee. This knowledge of examinee experience may be known (e.g., as in local assessments embedded in instruction) and employed in task design decisions in order to sharpen the evidentiary focus of a task. Conversely, this knowledge may be unknown (e.g., as in large-scale accountability tests) so that substantive knowledge about a model and capability to use it become confounded.

2.3 Some Relevant Results from Psychology

Norman (1993) distinguishes between experiential and reflective cognition:

> The experiential mode leads to a state in which we perceive and react to the events around us, efficiently and effortlessly. The reflective mode is that of comparison and contrast, of thought, of decision making. Both modes are essential to human performance (p. 15, 20).

Both modes of reasoning are involved in model-based reasoning. This section briefly notes some results from cognitive research that are useful for understanding and then assessing model-based reasoning.

2.3.1 Experiential aspects of model-based reasoning

An individual forming a model to comprehend a particular situation activates, assembles, and particularizes elements from long-term memory to create an instance of a model that is tailored to the task at hand. Kintsch's "construction-integration" (CI) model of text comprehension (Kintsch, 1988, 1994, 1998; van Dijk & Kintsch, 1983) provides insights into this process. Kintsch and Greeno (1985) apply the CI perspective to understanding arithmetic word problems. In their example, the models of interest are abstract Change, Combine, and Compare arithmetic schemas to which particular problem situations correspond.

For a simple word problem we can do in our head, model formation takes place in working memory, incorporating features of the situation from sensory memory and information from long-term memory. Features of the unique real-world situation activate elements of long-term memory, which in turn can activate other elements of long term memory or guide the search for new elements or patterns in the situation. The individual's goals regarding the activity can also influence what models are activiated. That is, conative (as well as affective; see Damasio, 1994) considerations play a role in activating elements of long-term memory and help determine which aspects of a situation and what level of detail will be addressed. This construction phase (the C in CI theory) is initiated by features of stimuli in the environment and activates associations from long-term memory whether they are relevant to the current circumstances or not.

A "situation model" emerges from the integration (the I in CI theory) of mutually reinforcing elements among immediate stimuli and retrieved patterns. The situation model constitutes the learner's understanding or comprehension of the situation in which particular elements of that real-world situation are synthesized with more generalized patterns or schemas from previous experience. Ideally, in the case of arithmetic and scientific models, the elements of appropriate formal models are activated, they correspond to elements of the real-world situation, and the situation is comprehended in terms of its salient elements through relationships in the scientific model (Larkin, 1983). By virtue of the associations in a scientific model among its narrative, representation, procedural, and strategic aspects, model formation sets the stage for further reasoning—surrogative reasoning—such as carrying out operations that would produce the answer to a simple problem or devising a plan to solve a multiple-step task. In the section on Model Formation, for example, we will refer to Duncan's (2006) study of domain-specific heuristics and explanatory schemas in molecular genetics.

The same cognitive processes take place when students reason with partial, incomplete, fragmentary, and intuitive building blocks rather than with correct scientific models (diSessa, 1993, refers to phenomenological primitives, or "p-prims"). A situation model will result based on patterns from the student's past experience, which together provide an understanding of the situation upon which to base further reasoning and action. Unlike the situation model of an expert, however, this understanding may be based on superficial features of the situation or misconceptions; for example, the "continuous push" p-prim is that an object will keep moving only if some force is continuously applied to it. Such understandings often suffice for everyday life. The "continuous push" p-prim does in fact give the right qualitative prediction that a box will stop moving when we stop pushing it, and this p-prim is a lot less complicated than a Newtonian explanation. But these understandings are not cast in terms of coherent conceptions that connect diverse situations and link them to effective general procedures and strategies. People reasoning in this way are employing model-based reasoning, to be sure, but not through the scientific models that are the targets of science instruction.

An individual's successful formation of a cognitive situation model around a scientific model requires not only the availability of the formal elements of the scientific model from long-term memory but the availability of the cues and patterns to activate it and relate its elements to unique situations (Redish, 2004). Experts have more information in long-term memory about models than do novices, but more importantly, they have more effective connections among them—including conditions of usefulness (Glaser, 1988). Experts' model formation processes are streamlined with extensive use to accommodate more rapid access, larger chunks, and routinized procedures ("long-term working memory," in Ericsson & Kintsch's 1995 terminology). For example, Chi, Feltovich, and Glaser (1981) asked novices and experts in physics to sort cards depicting mechanics problems into stacks of similar tasks. Novices grouped problems in terms of surface features such as pulleys and springs. Experts organized their stacks in terms of more fundamental principles such as equilibrium and Newton's Third Law, each stack containing a mixture of spring, pulley, and inclined plane tasks. The experts' categorizations suggest a well-practiced model formation process for comprehending real-world situations in terms of physical principles that are not apparent on the surface. Their situation models are linked, in turn, to mathematical representations for solving problems (Model Use) and evaluating the representation (Model Evaluation) and to strategies and procedures for carrying out these activities.

2.3.2 Reflective aspects of model–based reasoning

While scientific models can form the basis of an individual's thinking about a situation, they also are cultural tools that people can use to think and interact together. Model-based reasoning is thus a special case of what Wertsch (1998) calls mediated action. Seeing model-based reasoning as action underscores how science is not merely a matter of models, formulas, and procedures, but ways of thinking, talking, and acting in the world through patterns of knowledge and understanding that have been built up within a community of practice.

Processes analogous to those in the CI model take place in conscious model-based reasoning—among people, using tools and external representations, and over time spans of minutes, hours, or even years rather than milliseconds. Tools and external representations embody key relationships in a model in ways that enable computation and capture intermediate results to help circumvent the limitations of working memory (Markman, 1999). The activation of relevant information in long-term memory is echoed in literature searches and conversations with colleagues about problems they tackled in the past. The counterpart of refocusing one's momentary gaze is generating a scatterplot, looking for trends and outliers, and re-expressing residuals in a different format. The elements of a tailored, synthesized, and integrated model can be drawn from different domains and configured and reconfigured over the course of multiple drafts of a research paper. The correspondence between the elements of real-world situations and the entities in an instantiated scientific model may require repeated attempts to determine just what to address in the model, at what level of detail, and in what representational form to achieve the goals at hand (i.e., cycles of Model Formation, Model Evaluation, Model Elaboration, and Model Revision).

Managing one's own activities in the in their full complexity and extending over time requires being able to reflectively monitor one's progress, evaluate the effectiveness of work so far, keep track of where one is in the inquiry cycle, and determine next steps. Collectively these are metacognitive skills associated with model-based reasoning. White, Shimoda, and Frederiksen (1999) cited Piaget (1976)'s argument that being aware of and reflecting on one's cognition reflects an advanced stage of development and also Vygotsky's (1978) claim that children progress from relying on others to help regulate their cognition to being able to regulate it themselves. Section 11 draws on this work for the *design pattern* on assessing the coordination of aspects of model-based reasoning in the context of more encompassing activities. The key idea is providing scaffolding for monitoring and self-regulating activities so that students come to internalize them through experience.

3.0 Evidence-Centered Assessment Design

The *design patterns* described in this text support the authoring of tasks to assess students' capabilities to carry out the kinds of model-based reasoning sketched above, using the tools and concepts of an evidence-centered approach to assessment design (ECD; Mislevy, Steinberg, & Almond, 2003; Mislevy & Riconscente, 2006). Messick (1994) lays out the essential narrative of assessment design, saying that we

> …begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? (p.16).

Evidence-centered assessment design distinguishes layers at which activities and structures appear in the assessment enterprise, to the end of creating operational processes that instantiate a coherent assessment argument (as described later in this section). Table 2 summarizes the ECD layers. *Design patterns* are tools for work in the Domain Modeling layer, where research and experience about the domains and skills of interest that have been marshaled in Domain Analysis are organized in accordance with the form of assessment arguments.

In order to show how *design patterns* support this work, we extend Toulmin's (1958) general argument structure to the case of assessment arguments. By conceptualizing assessment as a form of argument, we can use *design patterns* as supports for design choices in terms of the elements of an assessment argument. For further discussion on how assessment arguments are then instantiated in the machinery of operational assessments—stimulus materials, scoring procedures, measurement models, delivery systems, and so on—the reader is referred to Steinberg, Almond, and Mislevy (2002), Mislevy, Steinberg, and Almond (2003) and Mislevy and Riconscente (2006).

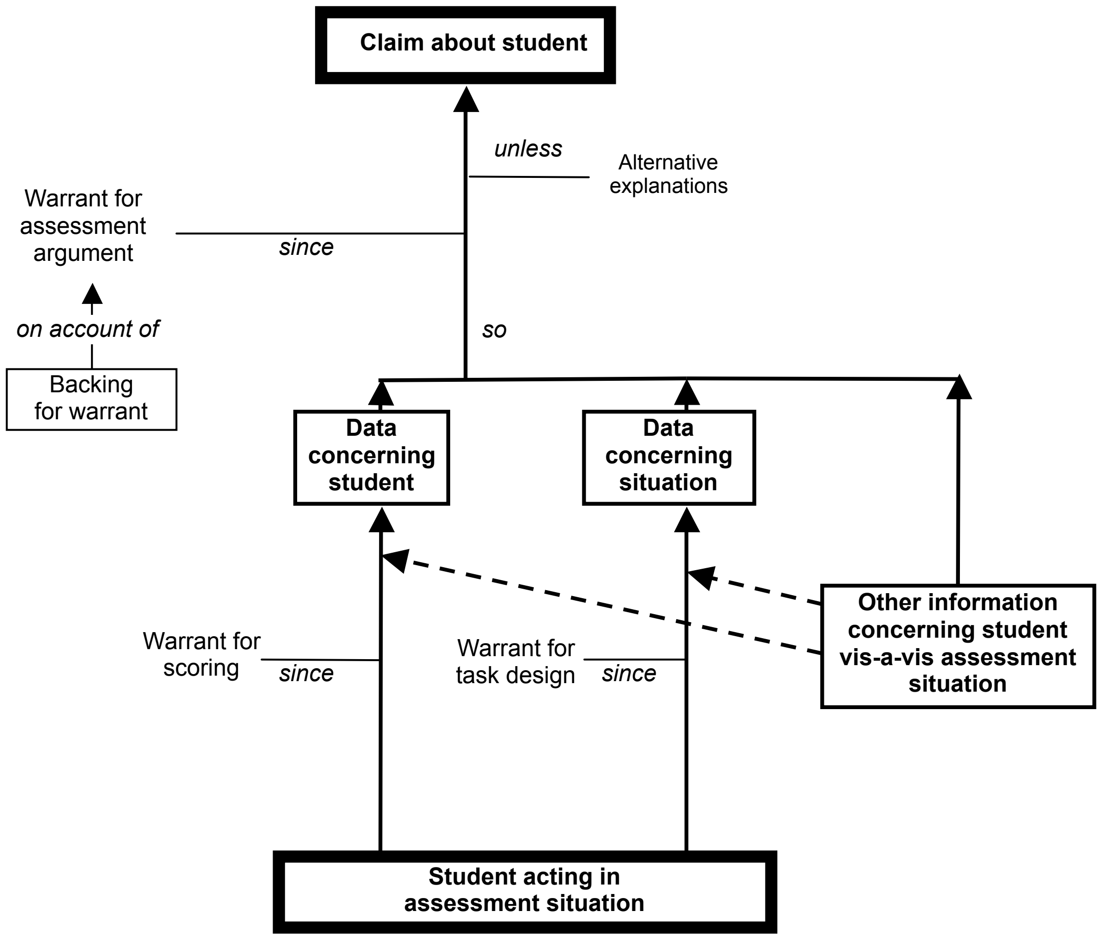
Table 2: Layers of Evidence-Centered Design for Assessments

Layer	Role	Key Entities
Domain Analysis	Gather substantive information about the domain of interest that has direct implications for assessment; how knowledge is constructed, acquired, used, and communicated	Domain concepts, terminology, tools, knowledge representations, analyses, situations of use, patterns of interaction
Domain Modeling	Express assessment argument in narrative form based on information from Domain Analysis	Knowledge, skills and abilities; characteristic and variable task features, potential work products, potential observations
Conceptual Assessment Framework	Express assessment argument in structures and specifications for tasks and tests, evaluation procedures, measurement models	Student, evidence, and task models; student, observable, and task variables; rubrics; measurement models; test assembly specifications; PADI templates and task specifications
Assessment Implementation	Implement assessment, including presentation-ready tasks and calibrated measurement models	Task materials (including all materials, tools, affordances); pilot test data to hone evaluation procedures and fit measurement models
Assessment Delivery	Coordinate interactions of students and tasks: task-and-test-level scoring; reporting	Tasks as presented; work products as created; scores as evaluated

3.1 Assessment Arguments

An evidentiary argument is constructed through a series of logically connected claims or propositions that are supported by data via warrants and are subject to alternative explanations (Toulmin, 1958). Figure 3 presents an evidentiary argument applied to educational assessment. The claims concern aspects of students' proficiency—what they know or can do in various settings. Data consist of their observed behaviors in particular task situations, the salient features of those tasks, and other relevant information the assessment user may have about the relationship between the student and the task situation, such as personal or instructional experience. Warrants posit how responses in situations with the noted features depend on proficiency. Some conception of knowledge and its acquisition—i.e., a psychological perspective—is the source of warrants and shapes the nature of claims a particular assessment is meant to support and the tasks and data needed to evidence them (Mislevy, 2003, 2006). In the present case, research on model-based reasoning grounds warrants; that is, it suggests how students with certain kinds of knowledge and capabilities for reasoning through particular models would be apt to do in what kinds of task situations. Alternative explanations for poor performance are deficits in the knowledge or skills that are required to carry out a task but are not focal to the claim, such as familiarity with the computer interface used in a simulation-based investigation—“construct irrelevant” requirements, in Messick’s (1989) terminology.

Figure 3: An Extended Toulmin Diagram for Assessment Arguments



3.2 Design Patterns

While Toulmin diagrams provide support for understanding the structure of an assessment argument, *design patterns* provide support for creating its substance. Table 3 lists the key attributes of a PADI *design pattern*, defines the attributes, and specifies which component of the assessment argument it concerns. *Design patterns* are intentionally broad and non-technical: “centered around some knowledge, skills, or abilities (KSAs), a *design pattern* is meant to offer a variety of approaches that can be used to get evidence about that knowledge or skill, organized in such a way as to lead toward the more technical work of designing particular tasks” (Mislevy & Riconscente, 2006, p. 72). Since *design patterns* do not include the technical specifics of domain content, psychometrics, or task delivery—these considerations come into play in the next layer of the design process, the Conceptual Assessment Framework

(CAF)—they provide a common planning space for the various experts that may be involved in the assessment design process, such as curriculum developers, item writers, psychometricians, teachers, and domain specialists.

Using *design patterns* to create assessment tasks provides benefits in terms of validity, generativity, and reusability. First, validity is strengthened as tasks inherit the backing and rationale of the *design patterns* from which they were generated. Creating a *design pattern* for some aspect of proficiency requires articulating the components of the assessment argument, including the line of reasoning that explicates why certain kinds of data can offer evidence about that proficiency. The *design pattern* is connected to backing, or the research and experience that ground, the argument. Laying out the argument frame before developing specific tasks in their particulars helps ground the interpretation of test scores. *Design patterns* remain a resource for subsequent task development, serving as explicit and sharable backing for new tasks in the same application or other applications that address the same areas.

A second benefit is generativity. Because *design patterns* organize experience across past research and projects that all address the assessment of some targeted aspects of learning, they support the creation of new tasks grounded in a strong line of reasoning. Organizing *design patterns* around aspects of learning, especially ones that are difficult to assess, helps a task designer get started much more quickly: Scaffolding is provided about the shape of the argument, approaches that have been used in the past, and examples of tasks that illustrate the ideas.

A third benefit of *design patterns* is reusability. A *design pattern* encapsulates key results of work from the Domain Analysis stage and reflects the form of an assessment argument. As such it helps to structure a test designer's work in both Domain Analysis and Domain Modeling. The same *design pattern* can motivate a great many tasks in different areas and at different levels of proficiency, all revolving around the same hard-to-measure aspects of, say, scientific inquiry; their particulars can be detailed with the content, purposes, constraints, and resources of the assessment at hand. Moreover, one *design pattern* can be a starting point for creating a new *design pattern* that is similar, more specific, or more general than the original *design pattern*.

Table 3: Basic Design Pattern Attributes, Definitions, and Corresponding Assessment Argument Components

Attribute	Definition	Assessment Argument Component
Name	Short name for the design pattern	
Summary	Brief description of the family of tasks implied by the design pattern	
Rationale	Nature of the KSA of interest and how it is manifest. Concisely articulates the theoretical connection between the data to be collected and the claims to be made.	Warrant
Focal Knowledge, Skills, and Abilities (KSAs)	The primary knowledge/skill/abilities targeted by this design pattern	Claim
Additional KSAs	Other knowledge/skills/abilities that may be required by tasks motivated by this design pattern.	Claim if relevant; Alternative Explanation if irrelevant
Potential Work Products	Things students say, do, or make that can provide evidence about the focal knowledge/skills/abilities.	Data concerning students' actions
Potential Observations	Features of work products that encapsulate evidence about focal KSAs	Data concerning students' actions
Characteristic Features	Aspects of assessment situations which are likely to evoke the desired evidence.	Data concerning situation
Variable Features	Aspects of assessment situations that can be varied in order to control difficulty or target emphasis on various aspects of KSAs.	Data concerning situation
Examples	Samples of tasks that instantiate this design pattern	
References	Research, applications, or experience relevant to task design under this design pattern	Backing

Additional attributes can include links to other *design patterns* that are related to the current *design pattern*, for example as special-case or part-of relationships.

4.0 Design Patterns for Model-Based Reasoning

Distinguishable aspects of model-based reasoning, involving different, though overlapping, kinds of knowledge and processes, must be coordinated in investigations. A task designer may want to concentrate on one or more selected aspects—such as building a model that fits a given set of data, or revising an existing model—or address all aspects together in more extensive inquiry. The designer may want to provide different amounts of scaffolding for different aspects of an investigation. The *design patterns* presented here therefore highlight distinct aspects of model-based reasoning in a way that supports either focused tasks (building on one or a few *design patterns*) or more extensive investigations (building jointly on several design patterns). The Appendix presents summary forms of the *design patterns*. They are discussed in the following sections and illustrated with some tasks that focus on a single aspect of model-based reasoning and others that address multiple aspects.

The aspects of model-based reasoning listed in Table 1 serve as the Focal KSAs of the *design patterns* present here. As noted, they are meant to guide task design across the range of scientific models which can differ in content and detail. Content and level of detail are therefore Variable Features of tasks in all of these *design patterns*, and familiarity with the content and representational forms associated with particular models is a corresponding Additional KSA of each *design pattern*. What will be common to all tasks motivated by a given *design pattern*, however, will be the Characteristic Features—those features that are essential in a problem setting if it is to evoke evidence about the Focal KSA. To assess Model Revision, for example, there must be an existing model, observations that are at odds with it, and a need to revise the model to accommodate the discordant information. On the other hand, such tasks may vary as to the scientific model of interest and other features such as whether

- the existing model was provided or generated by the student in earlier work;
- the task is focused solely on model revision, or model revision is a multiply-occurring aspect to be evaluated in the context of a larger investigation;
- students are working independently or in groups; and
- the students' work takes place in hands-on investigations, open-ended written responses, oral presentations, or multiple-choice tasks.

These possibilities are highlighted for the designer in the attributes Variable Task Features and Potential Work Products.

A key assumption shared by all the *design patterns* bears emphasis. These *design patterns* are constructed around aspects of reasoning, but model-based reasoning is always about something. These are general *design patterns* for creating specific tasks; that is, tasks that involve reasoning with particular models in particular circumstances. The terms, concepts, representational forms, and procedures associated with a model will always be intimately involved with tasks created from these *design patterns*.

Thus substantive knowledge of the model(s) at issue will be an Additional KSA in every *design pattern* that follows. This alerts the task designer to important design choices concerning the interplay among the model-based reasoning that is targeted by a task, knowledge of the elements and processes of the particular models, and knowledge of the substantive aspects of whatever situation is presented.

For example, if the desired focus is assessing students' capabilities at carrying out reasoning steps, as might be of interest when the target of instruction has been metacognitive aspects of model-based reasoning, it can be appropriate to use very simple models that are familiar to virtually all students. In this case, substantive knowledge as an Additional KSA is unlikely to be a source of construct irrelevant variance; that is, students are unlikely to perform poorly simply because they aren't familiar with the underlying model. Glaser and Baxter (1998) described such tasks as "open process, lean content." Alternatively, if the desired focus is assessing students' capabilities at carrying out reasoning steps with a particular model that is the focus of attention, as when that model has been a target of instruction, then the demands for knowledge of that model can be high. The assessor wants to know if the student can carry out reasoning with that model, and failure due to lack of familiarity and facility with that model is both possible and construct relevant. Glaser and Baxter (1998) described such tasks as "open process, rich content."

An important Variable Task Feature that applies to all *design patterns*, entails a number of Additional KSAs, and holds implications for decisions about Work Products and Observable Variables, is whether the task is to be carried out by a group a students or by students working independently. When tasks are carried out by a group, the Characteristic Features, Focal KSAs, Work Products, and so on concerning the targeted aspects of model-based reasoning are still pertinent. However, group tasks induce Additional KSAs concerning skills of communication, interaction, explanation, persuasion, and so on, that can also be targets of inference. To design such tasks, a task developer can draw upon multiple *design patterns* such as the Model Formation and "Participating in Collaborative Scientific Inquiry" (Mislevy, et al., 2003). The latter *design pattern* provides support for thinking about features of tasks, work products, and ways of evaluating performances when inferences about collaborative work are also needed.

Thus far, we have introduced ideas and issues that cut across all the *design patterns* treated in this paper. Although the first six *design patterns* that follow highlight specific aspects of model-based reasoning, their essential interaction in practice must not be neglected. This is the central concern of the final *design pattern*. To help keep this larger picture in mind, along the way we present some examples of focused *design patterns* that subsequently appear as components within a larger investigation.

5.0 Model Formation

A scientific model is a system of abstract entities, relationships, and processes. Every particular use of a model begins with the selection and assembly of particular elements from this model writ large, to establish a correspondence with particular circumstances—often real-world phenomena, but also possibly the entities, processes, and relationships in other models. This aspect of model-based reasoning is called Model Formation (although Model Instantiation would be apt as well). This section presents a *design pattern* for assessing model formation in isolation or as an integrated aspect of more extensive model-based reasoning.

The first column of cells in the table of *design patterns* in the Appendix summarizes a *design pattern* to support writing tasks to assess model formation. As discussed above, the attributes are organized according to elements of an assessment argument. They suggest design choices for a task developer to consider, jointly motivated by research on model-based reasoning and considerations of task design.

5.1 Rationale, Focal KSAs, and Characteristic Task Features

An important feature of the Model Formation *design pattern*, and of all those that follow, is that it doesn't specify any particular model, even though model formation is inherently about instantiating particular models in particular contexts. We are not proposing that model formation is a decontextualized ability, independent of particular models and contexts.[2] Rather, the Model Formation *design pattern* addresses those features of the contextualized processes of model formation that are similar across contexts and models—features that are sufficiently similar that we can offer assessment design support that can be useful across those contexts and models. In all instances, a task motivated by the Model Formation *design pattern* involves a real-world situation, such as a problem setting or corpus of data and a purpose that motivates forming a model. These are Characteristic Features of model formation tasks. Variable Features of model formation tasks will be discussed shortly. Three particularly important ones are whether the problem is provided or student-determined, whether potential models are provided or must be generated, and whether the model formation is an independent activity or part of a broader inquiry task.

The Focal KSAs of this *design pattern* are aspects of the model formation process as described above, all in the context of a given situation and model:

- Ability to relate elements of the model to elements of the situation, and vice versa.
- Ability to describe (i.e., narrate) the situation through the entities and relationships of the model.

[2] It is the case, however, that an individual can develop a generalized schema for the value and use of models, and procedures and strategies for using them, which can be called upon to guide reasoning with new models and in new contexts.

- Ability to pose relevant questions about the situation necessary to inform the construction of the model.
- Ability to identify which aspects of the situation(s) to address and which to omit, including the scope and grain-size of model.
- Decision-making regarding scope and grain-size of a model, as appropriate to the intended use of the model.

Depending on the purpose of the intended assessment, a designer may chose to focus on some aspects more than on others and to address them separately and specifically or as an ensemble. We will see in the Variable Task Features subsection how choosing particular features of tasks will tend to elicit one or another aspect of the model formation process. For now, we note that the first two aspects in the list highlight the correspondence between elements of the model and perceived features of the situation. The last three highlight the correspondence among the model, the situation, and the purpose of modeling. This entails identifying which aspects of the situation are relevant and which can be safely ignored, and justifying the degree of accuracy needed for the purpose at hand. The entries in the Potential Work Products attribute will show that these aspects of reasoning may be elicited explicitly, inferred from intermediate work or talk-aloud solutions, or only lie implicit in the student's formulated model.

5.2 Additional KSAs

Additional KSAs are other aspects of knowledge that may or may not be involved in a model formation task at the discretion of the task designer, in accordance with the context and intended use of the task. They call a task developer's attention to design choices that will intentionally elicit or minimize demands on particular models and on other knowledge, skills, and abilities. Primary among Additional KSAs—and essential to any model-based-reasoning task—is knowledge of the scientific model or models that will be involved in the task. The designer may, on the one hand, wish to assess students' ability to form models of a given type when it is known that the students are familiar with the terms and elements of the model. On the other hand, knowing both the elements of a model and being able to instantiate it in a given setting may both be of interest.

For example, Marshall (1993, 1995) asks students to select an arithmetic schema from the five they have been studying (Change, Group, Compare, Vary, and Restate) and then map elements of a word problem to its slots (Figures 4 and 5). A teacher using Marshall's curriculum is implicitly conditioning his or her inferences on the knowledge that the models and representational forms already are familiar to the students, in order to focus the evidentiary value of the task on model formation using these models. The use of the same tasks in a large-scale survey assessment would confound knowledge of the arithmetic schemas and representations with the ability to match them to real-world situations. This can be perfectly fine if the question is whether students can apply the model in question to problems like the one in the task, without sorting out sources of their difficulty when they cannot. Note also that both of these tasks intimately involve Marshall's particular representational forms for arithmetic schemas—an appropriate KSA to call upon for students studying word problems using these conceptual tools, but a potential source of construct irrelevant variance in a "drop-in-from-the-sky" assessment.

Figure 4: Task for Selecting an Appropriate Schema (Marshall, 1993, p. 167)

INSTRUCTIONS: Choose the one diagram below that fits this story problem. Move the arrow into the diagram you have selected and click the mouse button.

Dan Robinson recently drove 215 miles from San Diego to Santa Barbara to see his parents. When he arrived at his parents', he noticed that the odometer of his car registered 45631 miles. What was the odometer reading before he made the trip?

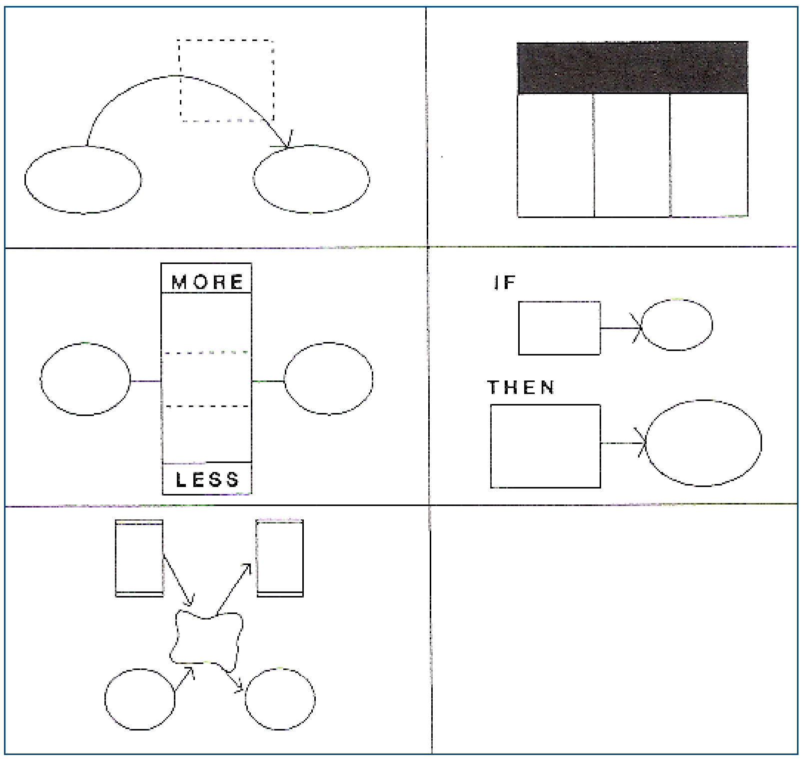


Figure 5: Task for Filling in a Schema (Marshall, 1993, p. 165)

INSTRUCTIONS: Identify the parts of the problem that belong in the diagram. Move the arrow over each part. Click and release the mouse button. Drag the dotted rectangle into the diagram, and click the mouse button again when you have positioned the rectangle correctly in the diagram. If you make a mistake, return to the problem and repeat the process. When you are finished, move the arrow into the OKAY box and click the mouse button.

Harry the computer programmer accidentally erased some of his computer programs while he was hurrying to finish work one Friday afternoon. Much to his dismay, when he returned to work on Monday, he discovered that only 24 programs of his original 92 programs had survived. How many computer programs had been destroyed?

As mentioned, knowledge of the model and/or content domain is always involved in model formation. Domain-specific knowledge structures, principles, procedures, and heuristics are at the heart of expert model formation. Newell and Simon (1972) called these “strong methods” for problem solving, in contrast to domain-independent “weak methods” such as means-ends analysis and trial and error. Seeing a physics problem in terms of Newton’s Third Law or a genetics problem in terms of a genes-code-for-proteins schema are examples of what Duncan calls domain-specific heuristics and domain-specific explanatory schemas. Being able to bring such strategies to bear will be important in any assessment where learning in the domain is at issue.

Familiarity with the task type and stimulus materials is another Additional KSA in model formation tasks and all other model-based reasoning tasks (and with any tasks in any assessment, we might add). For a student who is not familiar with a task type, irrelevant sources of difficulty can include what a problem is asking, how the information is presented, how responses are to be made, and expectations for responses. Because task features activate knowledge from long-term memory, the more often students have encountered a particular type of problem and addressed it with a given model, the more strongly the elements of that model will be activated by new problems that share features of familiar ones.

It is important for students to learn to solve so-called near transfer problems (Bransford & Schwartz, 1999) at an early stage of learning, and this is what some assessment tasks at this stage should address. However, unfamiliar tasks that yield to familiar models with novel mappings—i.e., far transfer problems—are more important in the long run (Clement, 2000). (And very far transfers, such as Louis de Broglie and Erwin Schrödinger’s formulation of a wave model for electrons in the early twentieth century, can be revolutionary discoveries.) It is by extending their experience across a range of situations with diverse surface features that students can begin to organize their knowledge in terms of underlying principles.

Interfaces, tools, representational forms, and symbol systems that appear in tasks can be essential to success, whether they appear as stimuli, are required in solution processes, or are needed to produce work products. A task designer interested in model formation with a given model will want to use only tools and representations students are familiar with in order to avoid construct irrelevant sources of difficulty. Although it is not a focus of this presentation, we note that other enabling knowledge and skills such as language, vision, and mobility that may be required in a task are also Additional KSAs, and will need to be minimized or circumvented to improve the accessibility of tasks for students with special needs (Hansen et al., 2005).

5.3 Variable Task Features

There is an important relationship in an assessment argument between Task Features, over which a task designer has considerable control, and Focal and Additional KSAs, which are aspects of the examinee’s capabilities (or lack therefore) that the task is meant to elicit. By making choices about Variable Task

Features, the task designer can include or exclude features that increase or decrease the demand for Focal and Additional KSAs. This should be done in a purposeful manner. There are particular relationships between Variable Task Features and KSAs that can be laid out to support a task designer in these design decisions.

One important Variable Feature was alluded to in the discussion of Focal KSAs—namely, the familiarity of the problem format and situation to model. Some familiar tasks may be useful for students first encountering a model. But unfamiliar tasks and contexts are necessary to assess students' capabilities to use given models in "far transfer" situations (Redish, 2004). Note that whether a task is "near transfer" or "far transfer" can generally not be determined by looking at the task. This depends on knowing the instructional and experiential history of a given student since the same task can be a familiar problem for one student, near transfer for another, and far transfer for a third. Data in the form of other information about the relationship between the student and the task (shown in the assessment argument diagram of Figure 3) play an important role in determining the value of this Variable Feature. If this relationship is not known (and it generally is not in a "drop-in-from-the-sky" test—aside from canonical examples such as inclined planes and Mendel's peas that are familiar to everyone who studies a domain) the evidentiary value of a student's response is degraded because of the alternative explanations that arise. Is a performance misleadingly good only because that particular problem was already familiar to a student? Is another student's performance on a usually easy task misleadingly poor because she had never seen that problem type before?

Another important set of Variable Features affects the difficulty of a task and can be used to create easy tasks for students who are first working with a model, very advanced and challenging tasks for advanced students, or tasks whose difficulty rests somewhere between these two extremes. Thus, the complexity of the model and situation to be modeled are two related, but distinguishable, kinds of features. Other things being equal, the need to use a more complex model makes a problem harder. Complexity features in a model include the number of variables or elements, the complexity of their interrelations, the number of representations required, and whether multiple models need to be used and integrated (also see the Model Elaboration *design pattern* on this last point). Complexity features in a situation include the number and variety of elements in the real-world situation, the presence of extraneous information, and the degree to which elements have been stylized in order to make their identification and model formation easier. As to the relationship between the model and the situation, difficulty can be increased by having more possible choices as to what to include in a model or how detailed to make a model in order to meet the goal of a task.

Tasks can vary in the degree to which students are familiar with the context in order to avoid extraneous knowledge requirements (as discussed in connection with Additional KSAs) or to intentionally incorporate requirements for substantive knowledge either because it is known that students are familiar with it or because that knowledge is itself a target of inference along with the capability to form models with it.

Tasks also can vary with regard to the amount of scaffolding provided. Marshall's schema selection tasks tend toward the more scaffolded end of the spectrum, as befits beginning work with a model.

Figure 6 is an example of a task with less scaffolding, developed by Patricia and Ken Heller at the University of Minnesota for use in cooperative group problem solving in physics. Their tasks

> are designed to encourage students to use an organized, logical problem-solving strategy instead of their novice, formula-driven, 'plug-and-chug' strategy. Specifically, the Minnesota group's context rich problems are designed to encourage students to (a) consider physics concepts in the context of real objects in the real world; (b) view problem-solving as a series of decisions; and (c) use the fundamental concepts of physics to qualitatively analyze a problem before the mathematical manipulation of formulas (K. Heller & P. Heller, 2001, p. 55).

Heller and Heller also reduce scaffolding by avoiding "trigger words" in their problem statements such as "starting from rest" and "inclined plane." These words activate physics schemas, to be sure—usually the correct ones in textbook exercises. But science educators want students to develop associations grounded in underlying principles rather than surface features of problem statements and to be able to form models in situations beyond stylized teaching examples.

Figure 6: Example of a "Context Rich" Problem (Heller & Heller, 2001, p. 104)

You've been hired as a technical consultant to the Minneapolis police department to design a radar detector-proof device that measures the speed of vehicles. (i.e. one that does not rely on sending out a radar signal that the car can detect.) You decide to employ the fact that a moving car emits a variety of characteristic sounds. Your idea is to make a very small and low device to be placed in the center of the road that will pick out a specific frequency emitted by the car as it approaches and then measure the change in that frequency as the car moves off in the other direction. The device will then send the initial and final frequencies to its microprocessor, and then use this data to compute the speed of the vehicle. You are currently in the process of writing a program for the chip in your new device. To complete the program, you need a formula that determines the speed of the car using the data received by the microprocessor. You may also include in your formula any physical constants that you might need. Because your reputation as a designer is on the line, you realize that you'll need to find ways to check the validity of your formula, even though it contains no numbers.

An example from the Performance Assessment Links in Science[3] (PALS) library further illustrates some of the Variable Features in model formation tasks. The Council of Chief State School Officers (CCSSO) contributed the task on the Predator/Prey relationship shown in Figure 7. Students are given a table with data regarding the population of hares and lynx in a particular area and are additionally told that the lynx is a predator of the hare. They are prompted to determine the relationship (i.e., formulate a model) between these two animals' population sizes. Tables and graphs are required, inducing the Additional KSA of familiarity with these representational forms. There is scaffolding for data analysis using a

[3] Downloaded July 31, 2007 from http://pals.sri.com/tasks/5-8/ME406/directs.html

coordinate graph—which ameliorates weaknesses with graphing techniques as an alternative explanation for poor performance—but no scaffolding for model formation. This combination of choices about what knowledge to support focuses the evidentiary value of the task on model formation rather than analytic methodology.

Figure 7: Predator/Prey Task

Population Dynamics: Predator/Prey Relationship

As a member of the International Committee for the Protection of Threatened and Endangered Animals (ICPTEA), you have been asked to respond to a subcommittee's report that there has been a rapid decline in the snowshoe hare population over the past four years. The major predator of the snowshoe hare is the lynx. In order to prevent the continued decline of the hare population, the subcommittee has proposed reducing the lynx population.

Previous research has shown that the snowshoe hare survives by eating the sparse plant material growing in the cold climate of Canada, and that the hare is capable of rapid population growth due to its high birthrate. The lynx has a much lower birthrate than the hare.

You have found the following data on the population levels of each species in a given region over a 28 year period (MacLulick, 1937). The population of hares is given in thousands, and the population of lynx is given in hundreds.

Time elapsed years	Population of snowshoe hare (thousands)	Population of lynx (hundreds)
0	20	10
2	55	15
4	65	55
6	95	60
8	55	20
10	5	15
12	15	10
14	50	60
16	75	60
18	20	10
20	25	5
22	50	25
24	70	40
26	30	25
28	15	5

To develop a clearer understanding of the research data in the table, plot the data on a line graph. Make sure that the axes are clearly labeled. Designate the snowshoe hare populations with a dot (.) and the lynx populations with an (x).

1. Using the data in the table and your graph, explain the relationship, if any, between the populations of lynx and snowshoe hares.
2. Write a response to the members of the subcommittee stating whether you support or reject the proposal to reduce the lynx population. Explain your decision using information you have obtained from the table of data and your graph.

5.4 Potential Work Products and Potential Observations

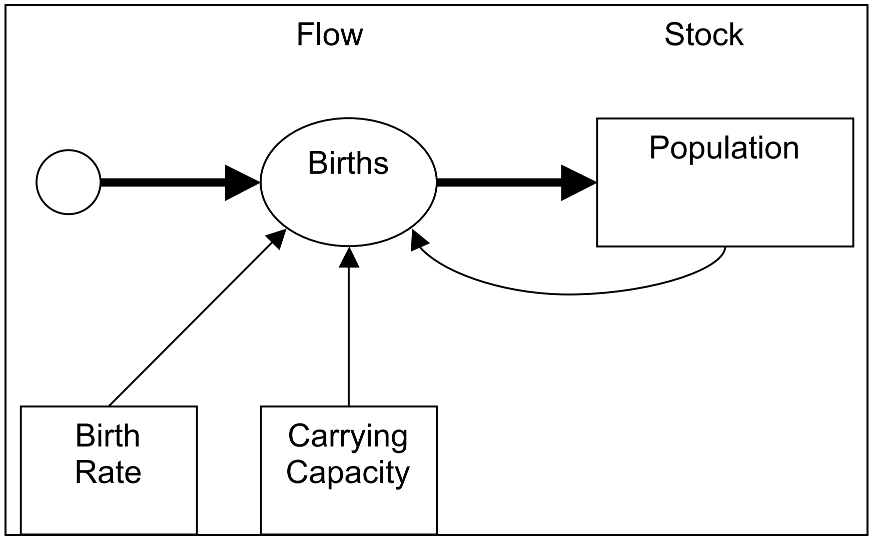
Because the cognitive process of model formation is not directly visible, an assessment argument must use for data the things students say, do, or make—the work products of an assessment task. Model formation tasks can be designed to elicit a variety of Potential Work Products, each varying in terms of its resource requirements, knowledge demands, the aspects of thinking it can provide evidence about, and the quality of the information obtained. A related design choice is determining which aspects of work products should be discerned and evaluated in a specific task. These are called the Observable Variables for the task, and they are evaluated from a student's performance. *Design patterns* provide support to a task developer by suggesting kinds of qualities that can be the basis for defining Observable Variables in a task. These suggestions are called Potential Observations. They provide various options for defining Observable Variables, some of which are available for use with a given work product or relationships among them. Potential Observations in a *design pattern* may be supplemented with rubrics, which broadly construed, are the processes—algorithms, instructions, or guidelines—which people or machines apply to Work Products to determine the values of Observable Variables that provide evidence about Focal KSAs.

The Potential Work Products attribute of the *design pattern* lists a variety of things students could say, do, or make to produce evidence about their model formation capabilities. A model formation task could produce Work Products associated with the final model that is generated, the process taken to produce it, and students' explanations and justifications of the model. The final model generally takes the form of one or more forms of knowledge representation, such as coordinated diagrams, a physical construction, or a system of equations with explanations of variables and relationships in terms of the real world situation. A Work Product could be the selection of a model from among a given set, such as Marshall's schema selection tasks; a constructed model in a constrained and therefore scaffolded work space, such as Marshall's fill-in-a-schema tasks; or a freely generated model in some representational form. With the availability of computer-based task administration, a wide variety of response forms can be used for students to express a model in constructive and open-ended ways that lend themselves to automated scoring (Scalise & Gifford, 2006; William, Mislevy, & Bejar, 2006).

When the form of the Work Product is produced with a technology-based tool, Additional KSAs are introduced with respect to both the familiarity with the representational form and use of whatever interfaces are required. On the other hand, use of such tools can be intimately related to understanding certain kinds of models, such as software programs used in interactive data analysis and modeling interactive systems. The Datadesk package for interactive statistical and graphical data analysis is an example of how software is integral to the targeted modeling knowledge and skill (Velleman, 1997). The STELLA package (Richmond, 2005) provides tools for students (and professionals) to build dynamic models, working back and forth among diagrams, equations, data, and graphs of interactive systems. Figure 8 is a stock and flow diagram similar to the ones the STELLA program uses for a simple model of population growth (Allen, Kling, & van der Pluijm, 2005), corresponding to the equation

$$\text{Births} = \text{Birth Rate}*\text{Population}*(1 - (\text{Population}/\text{Carrying Capacity})).$$

Figure 8: Stock and Flow Diagram for a Model of Population Growth



Work Products focused on the final product provide evidence about the quality of the model formation process, and hold clues as to which elements of the process succeeded and which did not. A student's STELLA model for the population growth model is an example of an unconstrained Work Product that contains clues about whether a student has dealt with feedback. For example, the constrained-format card-sorting physics task, (Chi, et al., 1979) yields as its Work Product the stacks a subject produces and the tasks in each stack, and contains clues about the characteristics of the problems the students used to group them.

Work Products that focus on the model formation process can include questions the students pose to themselves or to others, notes taken during the construction process, and traces (e.g., physical notes or diagrams, computer logs) of the steps taken during the formulation the model. These Work Products may be written or spoken; they may be captured on video tape, transcribed, or heard only by the assessor; they may be responses to explicit directives (e.g., answers to multiple-choice questions), answers to informal questions posed during instruction, or unstructured comments as obtained in talk-aloud solutions. Compared to final solutions, process-oriented Work Products can provide more direct evidence for metacognitive aspects of model formation and add support for instructional feedback.

Several kinds of Observable Variables can be evaluated from these various Work Products. Regarding final models, Potential Observations include the quality and accuracy of the final model, incorporating aspects such as the degree to which targeted aspects of the situation are represented in the model, the efficiency of the models and representations, whether extraneous elements are included in the model, and the appropriateness of the precision used for the goal of the task.

Kindfield's (1999) research on the use of diagrams to explain crossover in meiosis supports the value of using, as an Observable Variable, the inclusion of extraneous elements in a model: Novices' drawings were often more complete and better proportioned than experts', but what distinguished experts' diagrams was that only the salient features tended to be shown, and the relationships important to the problem at hand were rendered with whatever accuracy was needed to solve the problem. That is, the experts' diagrams were more efficacious than those of the novices.

When the Work Product is a functioning or runnable model, such as from STELLA, its behavior under the circumstances it is meant to approximate are potential features to evaluate. Does it capture the key elements? Does it function properly within certain ranges but fail outside others? Bearing in mind that many engineering approximations have this property, if the model fails outside a certain range, are the failures outside the scope of the real-world problem at hand?

Potential Observations regarding process can address time efficiency, the quality of self-monitoring questions asked, and the properties of intermediate models produced. For example, were there many restarts or scrapped work, as opposed to incremental improvements? How long was spent in planning before a first provisional solution was produced? For tasks involving mental models, rapid correct solutions as opposed to slow correct solutions provides evidence for automatized model formation, a characteristic of expert–like knowledge (Kalyuga, 2006).

When the Work Product is students' explanations of their models, Potential Observations include the quality and accuracy of the model-situation relationships, an awareness of considerations involved in choosing model elements, and degree of accuracy of modeling. In assessments where domain learning is at issue, Observable Variables based on whether or not (and if so how effectively) students employed the domain-specific heuristics and domain-specific explanatory schemas that are appropriate to the task.

In the genetics example introduced in Box 1, model formation is targeted when the students are given two sets of genes and must determine the probability for offspring to possess each type of trait (see Section 6.2 and Box 2). A model for the relationships is required to answer this question. If the Punnett Square has been previously introduced, the students must first simply recognize this as a situation in which the Punnett square can be used and consequently use it to explain how their models, in this form, fit the data. The students' model formation can be evaluated with regard to the accuracy of their instantiation of the Punnett Square in this situation. (The quality and accuracy of their answer segues into Model Use; Model Formation and Model Use are bound together in this task.)

Alternatively, suppose the students have not been exposed to this type of problem, and the instructor would like the students to formulate a model that is similar to a Punnett square in terms of the relationships among alleles and phenotypes. In this case, the instructor can present the students with information on different sets of parents' genes and traits and the genes and traits of their offspring. Using this information, the students must formulate how genes are combined and come up with the idea of the simple dominance model. The Work Product would be the representation of the final model, which can be evaluated for correctness and completeness.

Rubrics developed for model formation assessments provide the process by which features of Work Products are discerned and evaluated as evidence about the Focal KSAs. A rubric describes how features of one or more Work Products will be identified and expressed as a value of one or more Observable Variables. Observable Variables convey information on some aspects of the student's process or product in formulating a model in the performance of interest. Consider the two following examples.

In the PALS predator/prey task shown as Figure 7, students produce a graph of the relationship between the hares and the lynx and a written explanation for this relationship. These are the two Work Products. The rubric shown as Figure 9 assigns a score on a 1 to 4 scale, where a 1 signifies a wholly inadequate graph in terms of a list of targeted properties, and a 4 signifies a correct depiction of the relationship in a syntactically correct graph.

Figure 9: Rubric for Item 1 of the PALS Predator/Prey Task

Rubric	
NS	No attempt to graph (labels, numbers or plotting of any of the data onto the grid) is present. No adequate analysis (demonstration of understanding of the relationship between the snowshoe hare and lynx population by mentioning any of the aspects of the graph) is given.
1	Student demonstrates limited understanding of graphing and limited understanding of the relationship between the snowshoe hare and lynx populations. Example: An attempt to graph (labels, numbers or plotting of any of the data onto the grid) may be present. No adequate analysis (demonstration of understanding of the relationship between the snowshoe hare and lynx population by mentioning any aspects of the graph) is given. Student may indicate something about the data or, the trends or, the labels of their graph.
2	Student demonstrates some understanding of graphing and some knowledge of the relationship between the snowshoe hare and lynx populations. For example, the graph is constructed, the trends are accurate, and some data for the hare or the lynx is correctly plotted (but may be in thousands, not hundreds) or missing. The answer suggests that the student does not understand the relationship between the snowshoe hare and lynx populations OR the graph is not constructed correctly and plotted accurately, but the answer does demonstrate that the student understands the relationship between the snowshoe hare and lynx populations by mentioning at least one aspect of the graph.
3	Student demonstrates adequate understanding of graphing and adequate knowledge of the relationship between the snowshoe hare and lynx populations. Example: The graph is constructed correctly and data for the hare is plotted adequately (no more than three data points misplotted). Data for the lynx may be plotted in thousands, not hundreds, but has been adequately plotted (no more than three data points misplotted). An answer that demonstrates the understanding of the relationship between the snowshoe hare and lynx populations by mentioning at least two aspects of the graph is present (i.e., and increase in hare population leads to a lynx population increase; there is a delay in the change of the populations of snowshoe hare and lynx; there are ten times as many hares as lynx).
4	Student demonstrates a high level of understanding of graphing and a high level of knowledge of the relationship between the snowshoe hare and lynx populations. The graph is constructed correctly and data for the hare and lynx is plotted accurately. The difference in scale of the hare data and the lynx data is accurate. The correct analysis of the data is made by noting the three aspects of the graph. The delay in the change of the populations is noted (i.e., when the lynx population increases, years later the snowshoe hare population begins to decrease, and when the snowshoe hare population decreases, years later the lynx population begins to decrease).

In a study of students' learning in hypermedia environments, Azevedo and Cromley (2004) assessed the quality of the models students constructed to explain a diagram of the human circulatory system. The rubric, shown in Figure 10, summarizes a student's model in terms of increasingly more accurate and sophisticated understandings of the components and processes of the circulatory system. The Observable Variable is 'level of explanation,' as evaluated from a Work Product in the form of a transcript from a talk-aloud explanation. This rubric derives from research on progressive understandings of the circulatory system (e.g., Chi, 2005). The same backing could be used to create alternatives for multiple-choice items and to develop rubrics for working models of the circulatory system.

Figure 10: Necessary Features for Evaluating Models of the Circulatory System

Circulatory System Model – Rubric

1. **No understanding**
2. **Basic Global Concepts**
   - blood circulates
3. **Global Concepts with Purpose**
   - blood circulates
   - describes "purpose" - oxygen/nutrient transport
4. **Single Loop – Basic**
   - blood circulates
   - heart as pump
   - vessels (arteries/veins) transport
5. **Single Loop with Purpose**
   - blood circulates
   - heart as pump
   - vessels (arteries/veins) transport
   - describe "purpose" - oxygen/nutrient transport
6. **Single Loop - Advanced**
   - blood circulates
   - heart as pump
   - vessels (arteries/veins) transport
   - describe "purpose" – oxygen/nutrient transport
   - mentions one of the following: electrical system, transport functions of blood, details of blood cells
7. **Single Loop with Lungs**
   - blood circulates
   - heart as pump
   - vessels (arteries/veins) transport
   - mentions lungs as a "stop" along the way
   - describe "purpose" – oxygen/nutrient transport
8. **Single Loop with Lungs - Advanced**
   - blood circulates
   - heart as pump
   - vessels (arteries/veins) transport
   - mentions Lungs as a "stop" along the way
   - describe "purpose" – oxygen/nutrient transport
   - mentions one of the following: electrical system, transport functions of blood, details of blood cells
9. **Double Loop Concept**
   - blood circulates
   - heart as pump
   - vessels (arteries/veins) transport
   - describes "purpose" - oxygen/nutrient transport
   - mentions separate pulmonary and systemic systems
   - mentions importance of lungs
10. **Double Loop – Basic**
    - blood circulates
    - heart as pump
    - vessels (arteries/veins) transport
    - describe "purpose" - oxygen/nutrient transport
    - describes loop: heart - body - heart - lungs - heart
11. **Double Loop – Detailed**
    - blood circulates
    - heart as pump
    - vessels (arteries/veins) transport
    - describe "purpose" - oxygen/nutrient transport
    - describes loop: heart - body - heart - lungs –heart
    - structural details described: names vessels, describes flow through valves
12. **Double Loop - Advanced**
    - blood circulates
    - heart as pump
    - vessels (arteries/veins) transport
    - describe "purpose" - oxygen/nutrient transport
    - describes loop: heart - body - heart - lungs - heart
    - structural details described: names vessels, describes flow through valves
    - mentions one of the following: electrical system, transport functions of blood, details of blood cell

5.5 Considerations for Larger Investigations

The Model Formation *design pattern* is meant to support the authoring of both tasks that focus solely on model formation and tasks that include model formation as a part of a larger activity. Similarly, the next five *design patterns* (see Appendix) each target a specific aspect of model-based reasoning both for narrowly focused tasks and for tasks that encompass additional aspects. That larger context could entail model formation then model use, for example, or formation-use-evaluation, or a full investigation that engages all phases of inquiry. The full investigation could be scaffolded to distinguish model formation phases for the student, or the student could need to recognize when and how to form models. In the latter case, recognizing and managing phases calls upon the knowledge addressed in the final Model-Based Inquiry *design pattern*; the scaffolding in the former case supports this knowledge and thus does not provide evidence about them in order to enhance the value of evidence phase by phase. By determining features of the task situation in such ways, the task designer can tune the evidentiary value of the task to targeted aspects of model-based reasoning.

To make sense of extended performances in a larger task context, it can be useful to notice and evaluate model formation (as well as other aspects of model-based reasoning addressed in the following *design patterns*) as it takes place within that context. This is easier said than done in unstructured investigations. However, the points we have described regarding Characteristic Features, Potential Work Products, and Potential Observations still hold. They help the task developer evoke evidence about model formation in extended tasks, capture it in Work Products, and summarize it. Further, the task developer has a degree of control over how explicitly to elicit evidence about aspects of reasoning through design choices about which Work Products to require. In the design rationale for a simulation-based assessment of dental hygiene licensure candidates, Mislevy, Steinberg, Breyer, Almond, and Johnson (2002) found that the trace of actions—despite being a rich and detailed Work Product—did not convey students' intermediate mental products such as identification of cues, generation of hypotheses, and selection of tests to explore conjectures. They suggested introducing a Work Product in the form of an insurance form, similar to those now integral to the practice of dental hygiene. The student would need to indicate hypotheses based on cues from available forms of information about the patient and justify information-gathering actions with specific hypotheses or as standard-of-care for the situation.

5.6 Some Connections with Other Design Patterns

The Model Formation *design pattern* can be viewed as a subpart of the Model-Based Inquiry *design pattern*. To design and create extended performances in a larger context such as an inquiry investigation, the developer can use the Model-Based Inquiry *design pattern* to coordinate the overall activity and the constituent *design patterns* (such as Model Formation) to guide inquiry phases, Work Products, and Evaluation Procedures for the multiple phases of activity that will take place.

Many familiar tasks combine Model Formation with Model Use, the *design pattern* addressed next. A problem context is given, and a solution is required: The student must formulate a model and reason through it to obtain a solution. A task developer can choose among (1) evaluating the product of both aspects of model-based reasoning, so that evidence is evoked about either combined success or failure somewhere along the way, (2) obtaining discernable (though intimately dependent) evidence about formation and use by structuring Work Products that distinguish the stages, or (3) obtaining a rich Work Product such as a talk-aloud solution, traces of solution steps, or intermediate products and then seeking evidence by applying rubrics that address both the model formation and model using aspects of reasoning.

The Model Formation *design pattern* also overlaps with those for Model Elaboration and Model Revision. As an aspect of model-based reasoning, model elaboration focuses on combining or making additions to a model such as embedding it in a larger system or adding elements or submodels, or connecting to another model such that emphasis is placed on forming multiple, multilevel, or composite models. Model revision is a kind of model formation, but with a focus on responding to shortcomings from a given model as prompted by feedback from the environment, such as incorrect predictions or lack of fit to data.

It is possible to create finer-grained *design patterns* for model formation, such as having *design patterns* for mental models and *design patterns* for deliberative modeling. The *design pattern* presented in this section was meant to be broadly useful across domain areas, educational levels, and types of assessments. Thus, it necessarily offers less specific support for any particular area, level, or assessment type. More specialized *design patterns* could be developed in any of these respects, which would provide stronger support for task developers who need to develop tasks for certain needs.

6.0 Model Use

Model use is reasoning through the entities, relationships, and processes of a given model to provide explanations, make predictions, or fill in gaps with respect to real-world situations or summary data about real-world situations. Model use is a central aspect of model-based reasoning, and one would be reluctant to say that a student "knows" any model without being able to carry out reasoning of this kind. Model use is a necessary component in building, testing, and revising models. Some instruction and some assessment of model use with given models focuses mainly on reasoning through the relationships within a model, while other instruction and assessment will require model-use in coordination with model formation, testing, and revising.

6.1 Rationale, Focal KSAs, and Characteristic Task Features

Referring back to Figure 1, we see model use as making inferences about the real-world situation originally depicted at the lower left, through the relationships of the model in the middle plane—that is, in terms of the version of the situation reconceived through the eyes of the model, depicted on the lower right. Such thinking might be "run" in one's head as a mental model (Gentner & Stevens, 1983; Johnson-Laird, 1983) or supported by tools or external representations (e.g., a mechanical model or a computer simulation), as suggested by the links to representational forms and associated operations at the top of Figure 1.

It is useful to distinguish variations of this kind of reasoning that assessment tasks can highlight. Among the most important is *explanation* of a physical situation, which entails articulating the relationships among observations and events in terms of the underlying concepts, principles, and relationships of the model. For example, in order to give a "complete" explanation in a task from the Earth-Moon-Sun curriculum,

> …students have to put the relevant elements together into phenomenon-object-motion (POM) charts, which include an explanation using both text and diagrams, and articulate the relationship between their celestial motion model and the phenomenon in question (often using props such as inflatable globes, Styrofoam balls, and light sources) (Stewart, et al., 2005, p. 161).

Making predictions, constructing retrodictions (i.e., what might have happened previously for things to be as they are now?), and filling in missing information about a real-world situation are also varieties of model use, since one must reason through the relationships of the model to infer entities or circumstances in the future or the past, or that are otherwise not immediately observable.

It is useful to distinguish qualitative reasoning in terms of model concepts from the use of symbol systems and knowledge representations. Larkin (1983) showed that experts solved physics problems first by building an understanding of the situation in terms of the underlying principles and relationships—and only then proceeding to develop the systems of equations they eventually used to solve the problems. Hestenes (1987) argued that the emphasis placed on mathematical methods in college physics instruction and assessment slights conceptual understanding and biases students toward a formula-based approach rather than a model-based approach. Figure 11 is a typical task of the kind Hestenes deems insufficient. The formula-based approach may work fine for solving a set of problems at the end of a chapter, but it does not produce the desired deeper understanding of the underlying physics.

In response, Hestenes, Wells, and Swackhamer (1992) developed the Force Concept Inventory (FCI), a collection of tasks that present situations that require only qualitative reasoning through the fundamental concepts in kinematics. Figure 12 is an example from the FCI. Two similar assessments focused on qualitative reasoning through central models are the Force and Motion Conceptual Evaluation (FCME; Thornton & Sokolow, 1998), which addresses kinematics like the FCI but at a more advanced level, and the Test About Particles in a Gas (TAP; Novick & Nussbaum, 1981), which concerns the particulate nature and behavior of gases. We note that quantitative reasoning generally follows qualitative reasoning in practice, and cycling between the two is common. Looking ahead to a Variable Feature, model use assessment tasks can be cast to focus on just one or the other, or to elicit their sequential or cyclical interaction.

Figure 11: A Formula-Based Model Use Task

A projectile is fired horizontally from a flare gun located 45.0 m above the ground. The projectile's speed as it leaves the gun is 250 m/s.

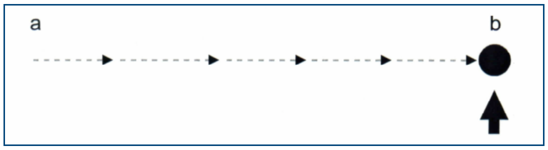
a) How long does the projectile remain in the air?
b) What horizontal distance does the projectile travel before striking the ground?
c) What is its speed as it strikes the ground?
d) If the projectile were simply dropped from a height of 45.0 m, instead of fired horizontally from that height, how much time would it take to reach the ground? How does this compare with your answer to part (a)?

Accessed on May 15, 2004 from http://ist-socrates.berkeley.edu:7521/projects/IPPS/Ch4/Prob6/Q.html, question 4-6.

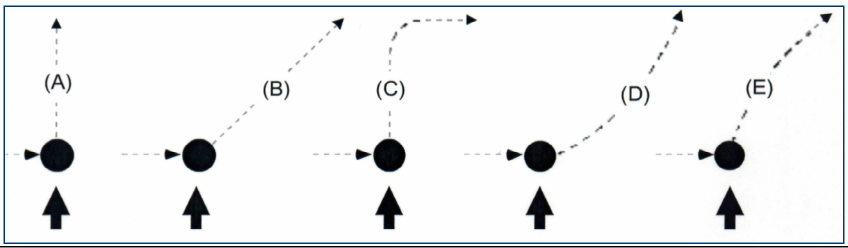
Figure 12: A Task from the Force Concept Inventory (from Hestenes, et al., 1992)

USE THE STATEMENT AND FIGURE BELOW TO ANSWER THE NEXT FOUR QUESTIONS (8 THROUGH 11).

The figure depicts a hockey puck sliding with constant speed v_o in a straight line from point “a” to point “b” on a frictionless horizontal surface. Forces exerted by the air are negligible. You are looking down on the puck. When the puck reaches point “b,” it receives a swift horizontal kick in the direction of the heavy print arrow. Had the puck been at rest at point “b,” then the kick would have set the puck in horizontal motion with a speed of v_k in the direction of the kick.



8. Which of the paths below would the puck most closely follow after receiving the kick?



The Focal KSAs at the heart of the Model Use *design pattern*, then, are the capabilities to make explanations, predictions, retrodictions, and fill in missing elements in the context of some model(s) and situation(s). This encompasses qualitative or quantitative manipulations, or both, as required.

All tasks based on the Model Use *design pattern* share Characteristic Features: a real-world situation and one or more models that the students will have to apply to this situation. For this aspect of model-based reasoning, the *design pattern* focuses on the fact that a model or models will be appropriate to the situation (at least provisionally, because model evaluation and model revision may need to follow) and that the student must reason through the models in order to reach some conclusion.

6.2 Additional KSAs

As with Model Formation, Additional KSAs that may be involved in a task for assessing model use include familiarity with the concepts, entities, relationships in a given model, and associated tools and representational forms. That is, both the declarative knowledge that is necessary to support reasoning through the model and whatever supports are required for apprehending, interacting with, and responding to a task must also be taken into account when drawing inferences from students’ performances. Demands for such ancillary skills can enhance a task’s evidentiary value, as when knowledge of representation software is known to be familiar to the examinee and can be used to support their reasoning—or it can degrade a task’s evidentiary value, as when examinees perform poorly due to a lack of necessary but ancillary capabilities (Wiley & Haertel, 1996).

Depending on the purpose of assessment, a user may be interested in all of these KSAs jointly or may chose to focus the evidentiary value of a task more selectively in light of what else is known about the relationship between the examinees and the task requirements. For example, an exercise may call for prediction from a model known to be familiar (solving another simple dominance problem at the beginning of a genetics unit), or solving a familiar kind of problem with a new model. Box 2, a continuation of the Genetics Toolkit Example, is an instance of the latter. Here students fill in a now-familiar Punnett square with regard to a co-dominance model just after the model has been introduced. As with other *design patterns* for model-based reasoning, model-using may be assessed in a task focusing on this aspect alone—model and data given, appropriateness presumed, at least provisionally—or as part of a larger task.

As noted in the previous section, tasks for assessing model use often require model formation. Model formation is an Additional KSA with respect to model use. A design choice that a task developer has, then, is whether to assess them jointly, separately, or sequentially. The following section includes a discussion of the task design tradeoffs that are entailed.

Box 2. Model-Based Reasoning Tasks in Genetics: Model Formation

Assessments of a students' proficiency in using models appear throughout Stewart and Hafner's genetics course. The same *design pattern* can be used for different assessments by modifying the model in question as well as other Variable Features. As the course progresses, the assessments may be more focused on other elements of model-based reasoning, but elements from model use are still involved. The following is a task that can be used at the beginning of the course when the focus is mainly on model use:

Complete the Punnett square to show the possible outcomes of a cross of a heterozygous father with a widow's peak with a homozygous mother with a widow's peak.

		Father's Genotype?	
		Possible Sperm?	Possible Sperm?
Mother's Genotype?	Possible egg?		
	Possible egg?		

A. What fraction of offspring would have a widow's peak?
B. What fraction of offspring would not have a widow's peak?

(http://www.cccoe.net/genetics/punnett4.html)

The model in this example is a co-dominance model for how alleles combine. Students are asked to reason through this model to apply it to make predictions regarding the offspring. Moreover, they must do so using the Punnett square representation, choosing the correct parent traits to cross and performing the crosses correctly. They then must be able to interpret the results in terms of possible traits of the offspring.

Notice that the answers that students give to problems A and B are dependent on them filling in the Punnett Square representational form appropriately. One possible Observable Variable is the joint correctness of the square and the question responses. A more nuanced rubric could first evaluate the correctness of the square and then evaluate students' question responses conditional on the way they completed the square. Even if they did not fill in the square correctly, they can still demonstrate some appropriate reasoning through the model by providing answers that are consistent with their square. For example, mistaking the relationship for simple dominance would lead to incorrect predictions, but reasoning from the Punnett square under this presumption does indicate appropriate steps of model use.

Providing the Punnett square is a design choice that supported students in using an appropriate tool for some steps in reasoning through the model. Not providing it would then provide evidence about whether a student could recall and use this representational form to reason through the inheritance model. Separate Observable Variables would be called for, as recalling the form is not equivalent to being able to reason through the model.

6.3 Variable Task Features

Variable Task Features for this aspect of model-based reasoning include, as with all other aspects, the model(s) at issue, students' familiarity with the type of model and the type of task, the complexity of the model(s) a student must work with, the amount and kinds of scaffolding that are provided, whether work is completed in a group or independently, and whether the targeted model use is embedded in a larger activity.

Connected with whether the model use is part of a larger activity are the questions of whether the data or the model are provided or generated by the student in a previous phase of a task. A design tradeoff arises: If model and data are provided, the developer can focus the evidentiary value of the task on whether the student can carry out the targeted reasoning through a model that is in focus. In this case, however, little information would be obtained about whether the student can manage the inquiry-cycle activities that characterize real-world model use. This decision may be appropriate when specified aspects of model use are the focus of instruction. Alternatively, suppose model use is assessed in a less structured manner, in which the student must collect or generate data, formulate a model, then use the model for further inference. Now difficulties in earlier stages of work may prevent the student from providing evidence about using the models of interest—even though evidence is obtained about a student's capability to manage the phases of inquiry (see the Model-Based Inquiry *design pattern*). A compromise design option is to stage an investigation in phases such that when students have trouble, say, forming an appropriate model, they are provided hints or scaffolding so that they can then carry out model use with the intended model.

Model use is carried out in the spirit of "as if"; that is, reasoning in the model space follows the rules embedded in the model for processes and relationships. An assessment task may be more difficult if students know that the model is incorrect in some way, even though this is exactly the kind of model use reasoning that will be required for model evaluation.

6.4 Potential Work Products and Potential Observations

Student Work Products that can be captured in model use include explanations, predictions, retrodictions, and filled-in information in the form of verbal, written, diagrammatic, symbolic, or physical media. Again, we refer the reader to Scalise and Gifford (2006) for taxonomy and illustrations of computer-based formats for Work Products that are amenable to automated scoring. Let the term "solution" stand generally for hypotheses, predictions, explanations, and/or missing elements of a real-world situation.

Three basic kinds of Work Product can be obtained to provide evidence about aspects of model use: The solution itself, traces of the solution, and explanations of the solution.

The solution itself. As in traditional large-scale assessments, this can take the form of the selection or identification of a solution from offered alternatives, as in multiple–choice, matching, true/false questions, and so on. Alternatively, the student may construct the solution in the form of one or more representational forms. This could be as simple as a word or number; it could take the form of a diagram or chart; it could involve a lengthy description of preconditions, possible causes of an event as reasoned from present evidence or predictions about possible outcomes and their likelihoods. The forms of solution may be generated by the student, or the student may fill in given, possibly partially completed, forms.

The FCI example in Figure 12 shows that a thoughtfully constructed multiple-choice task can provide a great deal of information about students' thinking. Like all FCI tasks, the distractors are designed to elicit common misconceptions about the domain. In this example, the curved options show paths that look very much like the parabolic paths of horizontally propelled objects that are subject to gravitational force—paths that are, in fact, correct answers to other FCI tasks that depict physically different situations. These distractors appeal to students whose understanding of forces is still at a surface level.

Traces of the solution. Traces of model using can be tracked as, for example, showing intermediate steps in problem-solving, capturing key stroke and action-selection in computer-based solutions, and talk-aloud protocols as the task proceeds. Martin and VanLehn's (1996) OLAE system for solving kinematics problems, for example, records each step of a student's solution, including restarts. These kinds of Work Products hold increasing value as tasks become more complex.

Explanations of the solution. A student can be asked to provide a written or oral description of a solution, how it was obtained, and its rationale. A presentation to other students is a formal and structured example. In contrast to the trace of a solution noted above, an explanation requires a student to verbalize steps, strategies, and rationales of the use of the model. The determination of the qualities of Work Products produces the data for inferences about student ability. Possible qualities to discern and evaluate, or Potential Observations, are the completeness and the accuracy of the reasoning of a prediction or explanation. They suggest ways that a task designer could define Observable Variables for a specific task for a particular purpose. The values of the Observable Variables would be determined from students' performances, as captured in the Work Products they produce, through the use of rubrics.

When the Work Product takes the form of a final solution, correctness and accuracy are usually of interest. In simple problems, this may suffice. In more complex problems, however, much thinking and many steps—hence, much potential information—takes place that may not be apparent in the solution alone. It can then be of interest to examine the steps taken in reasoning through the model and to evaluate the process in such terms as appropriateness, efficiency, systematicity, quality of strategy, and effectiveness of procedures. Evaluating Observable Variables such as the trace of a solution requires a method for detecting and summarizing its salient qualities. From a Work Product in the form of an explanation, the Observable Variables concern the student's capability to express these qualities. We note in passing that requiring explanations benefits instruction by making the steps of model use explicit and overt and hence, amenable to student reflection and supportive of metacognitive skills.

Choices about Observable Variables are linked to choices about Work Products. That is, there may be some Potential Work Products that support a given Potential Observation and others that do not, and if a given observation is desired, a Work Product that can provide it must be used. If evidence about students' capabilities to build a diagrammatic representation from a verbal description is desired, then a labeled diagram is an appropriate Work Product to elicit, and Observable Variables pertaining to its adequacy and correctness are called for. If evidence is desired about associating elements of equations to aspects of diagrams, students can be asked to generate or select equations, and rubrics are needed to evaluate adequacy and correctness. If evidence about misconceptions is desired, a number of different Work Products could be elicited, as long as both the construction of the task and the evaluation procedures made it possible to evoke a misconception and note evidence about it: Rubrics can be used to identify and characterize misconceptions from open-ended Work Products such as explanations or written solutions from which they must be identified, while closed-ended Work Products such as multiple-choice responses on the FCI present options that reflect particular misconceptions.

When the Work Products in a given task include explanations, it is also possible to evaluate the quality of students' reasoning about their own reasoning. This kind of Work Product and Observable Variable draws attention to metacognition—specifically, the quality with which student's are monitoring and evaluating their own use of the model. Especially when students are asked to evaluate their own or other students' reasoning as they use models, a critical awareness of the way one is reasoning is an important way to develop model use capabilities.

6.5 Some Connections with Other Design Patterns

As noted above, tasks that combine model formation and model use are common: A student is presented a real-world situation and asked to solve a problem, provide an explanation, or make a prediction or retrodiction. In simple problems, these aspects of model-based reasoning are difficult to individuate. In FCI multiple-choice tasks, for example, the only Work Product is the response choice, in which the distractors have been constructed so a correct response suggests the student both formulated and reasoned through the correct model, while an incorrect choice suggests the student has formulated and reasoned through a model based on a particular misconception (see Cromley & Mislevy, 2004, for further discussion of *design patterns* and *task templates* for tasks based on misconceptions).

It is possible, however, to capture evidence separately for model formation and model use in tasks that require both aspects of reasoning. This is done by requiring Work Products that specifically express the model being formed, then the reasoning through that model. Observable Variables based on the multiple Work Products—each motivated by the corresponding *design pattern*—can then be developed to distinguish phases of model-based reasoning.

Most tasks that address model evaluation and model revision also involve model use. In both cases, it is necessary to reason through a provisional model in order to compare its predictions with a situation. For model evaluation, the model in focus is the model being evaluated. In model revision, the models being reasoned through are alternatives to a given model, to see if their predictions accord with the situation better than the present model. *Design patterns* for these aspects of model-based reasoning will be discussed below.

Reasoning within the model space using symbol systems and representations is important and merits its own *design patterns*. Interest in this presentation focuses on how manipulations and relationships within the model representation correspond, or do not correspond, to relationships in the real world. One particular aspect of working between representations and models is addressed, however, in the model articulation *design pattern*.

7.0 Model Elaboration

In real-world scientific practice, new data sometime require dramatic revisions to our current scientific understanding. Much of the time, however, new data extend a current model or integrates multiple portions of familiar models. This is called model elaboration. In addition to being a frequently-engaged aspect of scientific activity, model elaboration constitutes an important part of the learning process, as it involves students in making connections across elements of their content knowledge and deepening their understanding of scientific theory. Model elaboration is closely tied to the structure of scientific theories themselves, which, according to Giere (2004), can be viewed as populations of models which can be assembled to reason about simple or complex situations in their scope. Frederiksen and White (1998), for example, demonstrated a module for learning about electricity that consisted of a sequence of increasingly elaborated models.

7.1 Rationale, Focal KSAs, and Characteristic Task Features

Stewart and Hafner (1991) identify four ways that engaging in model elaboration benefits students' ability to reason scientifically.

1. *Learning more efficient procedures for generating data.* In genetics, simple dominance problems can be solved by crossing individuals with unlike variations for each trait to identify heterozygotes (thus, the dominant variation).

2. *Developing within-model conceptual insights.* These can take many forms, such as when an individual extends specific relationships to arrive at more general characteristics of a physical phenomenon.

   > For example, in multiple alleles more than two alleles of a gene exist in a population, but no more than two can occur in an individual. One elaboration of this model is the recognition that the number of variations for a trait is a function of the number of alleles in the population and the types of interactions (simple dominant or codominant) among pairs of alleles (Stewart & Hafner, 1991, p. 113).

3. *Linking models because they share objects, processes, or states.* This involves generalizing from special cases.

   > An example of this is linking models of simple dominance and meiosis. This linking provides the physical basis of chromosome behavior for understanding segregation and independent assortment, both theoretical constructs in the simple dominance model. A second example of linking is the realization that an individual can only have two alleles (and that the two alleles can interact in a simple dominant or co-dominant fashion) no matter how many alleles exist at a locus within the population. Thus, multiple alleles can be seen as a special case of simple dominance and co–dominance, or conversely, these two can be viewed as simple cases of multiple alleles (Stewart & Hafner, 1991, p. 113-114).

4. *Linking models to produce a larger framework.* This entails development of overarching principles that traverse a wide range of problems. An example would be a "big picture" of genetics developing from solving a wide range of problems.

Passmore and Stewart (2002) describe a classroom curriculum that includes student elaboration of the Darwinian model:

> Once students had initial experiences composing Darwinian explanations and had explicitly considered the components of an appropriate explanation, they were given a data-rich case from which they were expected to develop a more complete Darwinian explanation. This case was designed to provide students with an opportunity to investigate a change in a trait over time, to use the natural selection model to explain that changes, and to support their argument with appropriate pieces of evidence. We intended to create a setting in which it was necessary for students to deepen their understanding of the components of the natural selection model in the course of using those components to create an explanation for the case phenomenon (p. 194).

As illustrated in this example, if students are to develop this quality of knowledge, they must be exposed to the kinds of experiences that afford this extending and restructuring of their understandings of the scientific models under investigation. Assessment tasks that address model elaboration will extend a model or address interconnections between or within models. Task situations typically present a situation in which familiar or currently-targeted models are required, but combinations or connections among them are required to formulate a model for the situation. Connections across individual-level and species-level models in evolution and between quantum and classical models in physics illustrate opportunities for learning and for assessing model elaboration.

A simple example of a model elaboration task can be obtained by nesting arithmetic schemas in a multi-step problem. Extending the single-schema task in Figure 4 yields the two-schema situation shown as Figure 13.

Figure 13: A Task Requiring the Nesting of Two Arithmetic Models

> Klaus Frisch recently drove his American-made automobile 265 *kilometers* from San Diego to Santa Barbara to see his parents. When he arrived at his parents' house, he noticed that the odometer of his car registered 45631 *miles*. What was the odometer reading *in miles* before he made the trip? (Hint: 1 kilometer = .6 miles)

Whereas the original problem required forming a reconception of the situation in terms of the Change schema, this two-step problem first requires the formation of a Vary schema to translate kilometers traveled into miles traveled, which as a composite fills in a slot in a Change schema. As a Work Product that emphasizes the relationships among schemas in multi-step problems, the student could be asked to drag the representation of one schema into another, then fill in the slots of the more complex assembled representation with the information given in the problem statement.

Tasks that elicit model elaboration possess several Characteristic Features. Tasks should be set in real-world situations for which an elaborated model is required, in terms of requiring linkages between models or extensions of the elements of a given model. Note that this characteristic is only fully understood in light of the students' history, in that extending a model in a given context may be a new experience to one student, but to another student involve simply applying a familiar model. If constraining students to model elaboration rather than model revision is desired, the situation or data should be compatible with the models accessible to students. The task solution must involve combining or making additions to existing models. Examples include embedding a model in a larger system, adding more parts to the model, or incorporating additional information about a real-world situation into the schema the model represents that in some way modifies the modeled representation.

7.2 Additional KSAs

Because model elaboration regards the structure of knowledge, Additional KSAs concerning subject-matter knowledge are of particular importance. Content knowledge is a prerequisite to model elaboration. Thus, failure on a model elaboration task can be due to lack of subject-matter knowledge. Only if we can rule out lack of subject-matter knowledge as an explanation for poor performance can we infer troubles with model elaboration in the given task context. Note, for example, that the hint in the two-schema odometer problem provides the relationship between miles and kilometers in order to remove failing to know this fact as an alternative explanation for poor performance (as opposed to the targeted model elaboration).

As usual, familiarity with task expectations, materials, and procedures are Additional KSAs that enable or hinder performance and must be taken dealt with by design choices for materials, Work Products, and evaluation procedures in light of the testing population and context.

7.3 Variable Task Features

The substance and particular models involved in a model elaboration task are central Variable Features of tasks. Learning tasks will often build on models that students are already familiar with in order to maximize opportunities to further the students' understanding. The Genetics Toolkit example discussed in Box 3 is such an example.

A model elaboration task can address elaborating or extending a given model, or connecting multiple models. One could split the model elaboration into two more narrowly defined *design patterns* along this distinction.

As with other *design patterns* in this collection, Variable Task Features include whether the task provides the data or situation that is the object of modeling, whether the aspect is the sole focus of the task as opposed to being part of a larger activity, whether the task is to be addressed by an individual or a group, and how much or what kinds of support to provide. One kind of support concerns the model(s) that are the focus of elaboration: Are hints or direct instructions offered for the model(s) to be elaborated, or are they to be provided, unprompted, by the student? Supports can also be used to reduce or circumvent the demand on construct-irrelevant Additional KSAs.

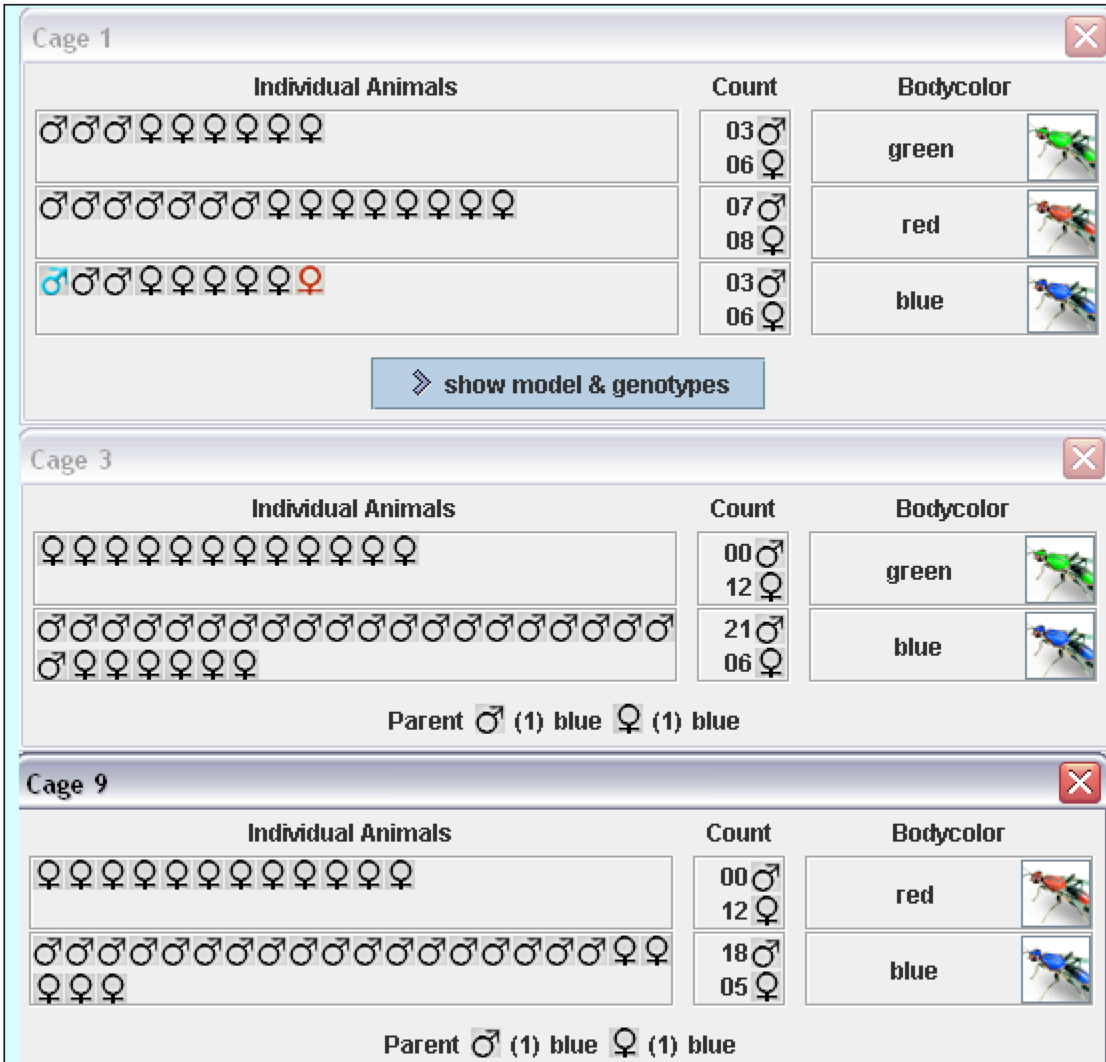
Use of knowledge representations and tools is also a Variable Feature of tasks. Involving a representation or tool can be a support for students who are familiar with it and can increase the evidentiary value of the task. On the other hand, if when the assessor does not know whether students are familiar with representations or tools, requiring their use introduces an alternative explanation of poor performance and degrades evidence about Focal KSAs.

The degree and complexity of elaboration is a Variable Feature. Is a straightforward elaboration of a familiar model required, or less obvious extensions within or across models? Or are multiple models involved?

Box 3. Model-Based Reasoning Tasks in Genetics: Model Elaboration

It is common for students to first learn a simple model, then learn to extend it to incorporate more variables or additional situations. Model elaboration can be assessed by presenting a student with a familiar model and additional information that requires extensions of the original model to accommodate the new information.

In genetics, students generally learn about the simple dominance model first. They will then be given problems that may ask them to determine the possible outcomes of a cross, or based on the outcomes of a cross, to identify the dominant and the recessive traits. Students may then be given the information that for some traits there are more than two alleles. The Virtual Genetics Lab presents situations in which there are three alleles for the color of a bug. In this case the possible colors are blue, green, and red; the possible alleles are represented as A, B, C, where A is dominant to B and C and will lead to a blue bug, B is dominant to C and will lead to a green bug, and C is recessive to both A and B and will lead to a red bug. In this lab students are given possible bugs and are asked to cross them in order to determine which traits are recessive.



The simple dominance model that the students had previously learned is a model for alleles and rules for their combinations. It fits a number of real-world situations. In this task, students must extend their understanding of the rules in the simple dominance model with two alleles to situations in which there are more than two alleles, combining under an extended set of analogous rules. They must realize that each pair will have a dominant and recessive trait, and while one allele may be dominant in one situation it can be recessive in another. This elaboration extends their inheritance model to more real-world situations. For this task, familiarity with the simple dominance model is an Additional KSA. In a class where the teacher knows the students are familiar with the simple dominance model, this knowledge can be presumed. In a large-scale test, the test designer could choose to provide the simple dominance model in order to better focus the evidence on model elaboration.

This basic task could be varied with respect to the trait in question, the number of alleles, and the types of relationships in the model. For example, is trait dominance strictly ordered or are there instances of circular dominance? Different choices can increase or decrease the difficulty of the task, while still providing evidence about the students' capabilities with model elaboration. For this problem the students are only asked to select organisms to cross, and then click a button to obtain the results. Variants of the task could require more complicated procedures for students to test out their model, such as requesting cross of offspring from an initial set of crosses.

Work products could include the students' answers about which alleles are dominant to other alleles, in open-ended or multiple-choice forms; the crosses that they performed; and explanations of their reasoning. With regard to Potential Observations, instructors could characterize how systematic a student's choices of crosses were. From explanations, they could determine if a correct answer was based on appropriate or inappropriate reasoning, through a correct or in some way flawed model.

7.4 Potential Work Products and Potential Observations

From tasks with Characteristic Features of the model elaboration *design pattern*, students generate the Work Products that may include representations of their elaborated models (including, for example, nested representations of model schemas as in Marshall's SPS or STELLA models that incorporate familiar sub-models), oral or written explanations, traces of their steps while developing their elaborated models, and mappings of their elaborated model with a real-world situation.

Several Potential Observations can be identified with these Work Products, such as the accuracy and completeness of the linkages in students' elaborated models, the extent to which the elaborated model is accurately linked to the real-world situation, and the quality of explanations students provide for their finished product and the path they took to get there. Of particular importance as Potential Observations in model elaboration are (1) appropriateness in the region in the model space where the extensions or connections are required, and (2) appropriateness of the correspondence between the modeled situation and the posited model in the region in which the elaboration is required.

7.5 Some Connections with Other Design Patterns

Model elaboration can be considered a special case of model formation, in that the aim is to develop a modeled conception of a situation (then perhaps carry out further reasoning with it). However the emphasis in model elaboration is on what is happening in the model layer with respect to extensions of models or connections between models, even as these may be motivated by the real-world situation.

Model elaboration is also similar to model revision, in that a given model or a set of unconnected models does not account properly for the target situation and reformulation is required. It differs by its more particular focus on extensions and combinations of models rather than modifications within a given model's aegis, to rectify discrepancies in the model/data correspondence. Again, the point of particular interest in model-data correspondence is the areas that require extensions of a given model or the coordination between models.

8.0 Model Articulation

As conceptual knowledge structures containing content, procedural, and strategic information, scientific models admit to representation in a great variety of forms. They are indicated by the layers on the highest plane in Figure 1. Examples include force diagrams in physics, Punnett squares in genetics, and algebraic equations. These representations can take quantitative or qualitative forms, and multiple representations often exist within the scope of a single scientific model. For instance, both force diagrams and algebraic equations can be used to express Newton's laws of motion. In the domain of probability, both mathematical equations and path diagrams linking latent variables are employed in structural equation modeling (SEM). These representational systems allow us to reason in different ways about different aspects of a given scientific model and the real-world situations it signifies.

Although these representations vary greatly in their form, they share a symbolic nature: whether written characters or objects, visual or spoken, knowledge representations are expressed with symbols. Circles, squares and arrows are used to illustrate structural equations diagrams; alphanumeric characters and operator symbols are the building materials for mathematical equations. A symbol system encompasses a set of symbols, interrelationships among symbols, and valid operations for acting on symbols. The constitutive markings, notations, or sounds of symbol systems are distinguished from the meanings they denote (Smith, 1983; Greeno, 1989). A scientific model with one or more such representations can be conceived as a model system comprised of objects or entities in the model (e.g., model genes, model particles, model molecules) and the relationships and processes that characterize them (e.g., modeled mutation, modeled atomic structure). The relationship between the qualitative entities and relationships of the model or abstract systems establishes the meaning of symbols and operations in the symbol system.

8.1 Rationale, Focal KSAs, and Characteristic Task Features

The importance of the ability to navigate from one system to another in science is well-illustrated in findings such as Larkin's (1983) research on physicists' reasoning processes. When presented with a force problem, the expert physicists first took a qualitative approach, identifying salient entities and their interrelationships and singling out appropriate models for solving the problem (e.g., "This is an equilibrium problem."). From their resulting understanding at the qualitative or narrative level, they proceeded to build a set of equations (i.e., a symbol system) that corresponded to the situation and to solve the problem by working through the equations. By connecting the physical situation to a description in the semantic terms of the model and then in turn to the symbol system, the work carried out within the symbol system acquired a situated meaning and responded to the force problem.

As Greeno (1989) notes, however, much of the learning that takes place in classrooms targets development of students' ability to reason within a particular (typically symbolic) representational system.

Examples include fluency with the symbolic notations, operations and relationships of linear algebra. While these are necessary skills for reasoning with scientific models, simply being able to carry out manipulations strictly within a symbol system layer is not sufficient. Ability to reason between systems is an essential aspect of scientific inquiry and thus represents an important target for instruction and assessment. This includes translating meanings between the semantic system of a model and an associated symbolic system, or from one symbolic system to another within the context of the model. For instance, force diagrams and algebraic equations are two symbolic systems for Newton's model of motion. Evidence of students' ability to accurately translate force diagrams into mathematical equations (and vice versa) support claims about a student's capability to reason appropriately with physics models for force and motion.

The Model Articulation *design pattern* (see fourth column in Appendix table) supports developing tasks to assess articulating meanings between systems associated with a model. Focal KSAs concern making the connections, translations, or re-representations of information within a model system, across representational systems associated with the system. This includes the mappings between the conceptual or semantic entities, relationships, and processes within the model (the middle layer of Figure 1) and formal representations (the upper layer). It also includes expression across formal knowledge representational systems, such as graphs and mathematical expressions that are more widely applicable, as contextualized within a given substantive model.

Characteristic Features of tasks that assess model articulation are the involvement of multiple representation systems, and the need to translate meaning or information across these forms. This may take various forms, such as semantic formulation and an associated mathematical formulation, or a semantic formulation and a physical model, or two different symbolic representations such as mathematical expression and graphs within the context specified by the model of interest.

Model articulation differs from model formation in that it focuses on reasoning at the semantic layer and associated representational layers above it rather than the correspondence between a model and a real-world situation. However, when model articulation is addressed within the context of a real-world situation, the situation imposes constraints on connections among representations and, in some cases, provides situated meanings for connections among representations.

8.2 Additional KSAs

As usual, Additional KSAs that can arise in this *design pattern* are content knowledge and familiarity with task expectations, materials, and procedures. More particularly, inferences made about model articulation ideally would proceed under the assumption that students already are equipped with the capability to reason *within* a given system, since failing to do so would provide an alternative explanation for task failure. Mapping between force diagrams and algebraic representation in mechanics, for example, can fail if a student is insufficiently skilled with the calculus needed to express a targeted relationship. Therefore, while the Focal KSA in this *design pattern* targets ability to articulate *between* systems, knowledge *within* systems serves as an Additional KSA.

8.3 Variable Task Features

Which model system is addressed is as always a Variable Feature of tasks. Within this selection comes the choice of which particular representations to address. A key distinction is this: Is the targeted articulation between (a) the semantic layer of a model and a representational system, or (b) representational systems within the context of the model system?

Tasks motivated by this *design pattern* can vary in the number and combinations of systems included. Some tasks may include only a single symbol system and a single model system and ask students to describe the symbols in terms of the model, or vice versa. Alternatively, tasks may require students to consider two symbol systems associated with the same scientific model and to re-express the meaning of an expression in the first system with an expression in the second. In these cases, Potential Observations would include the accuracy and completeness of students' mappings from one system to another.

Other tasks may ask students to express a prediction in terms of one system based on a given representation in a second system. For example, "What will happen to the velocity of the ball described by equation b?" calls for articulation between the mathematical representation and the semantics of the Newtonian model. Physical representations can also be used, as in basic math when elementary students are given a subtraction problem and told to implement it with physical objects (e.g., 6-2 = ? can be represented as removing two blocks from a pile of six blocks).

The complexity of the systems and mappings are variable as well. Requiring transformations within systems as well as across systems increases task difficulty, although it does increase the requirements for the Additional KSAs for capabilities within those representational systems. Not only does this call upon within-system operations, it requires greater understanding of the set of relationships among all the components of the model system. This may be construct-relevant and appropriate in some contexts, and therefore appropriate as is — or it may be preferable to scaffold within-system operations in order to sharpen the focus on the articulation between systems.

Another central Variable Feature in this *design pattern* is whether the articulation in focus is prompted. On one hand, a task designer can explicitly call for a mapping or interpretation between the semantic and a specified symbolic system, or between two specified symbol systems. On the other hand, evidence for articulation between systems may be sought in an open-ended solution to a problem or during the course of an investigation without prompting. In this case, the Work Products may or may not contain evidence, and if they don't, it will not be possible to evaluate the Observable Variables. Unprompted evaluation of model articulation is necessary when a task is meant to assess the student's recognition of the need and appropriateness of alternative expressions within the model space.

8.4 Potential Work Products and Potential Observations

The main Work Products that convey evidence about model articulation are (a) re-expressions of information about elements or relationships within a model system across multiple representations, and (b) explanations of such representations. These could be in closed form, as with multiple–choice tasks in which alternatives offered different re-expressions or explanations, constructions of representations either from scratch or in partially completed forms, or a trace of activities leading to articulation across representational systems. As noted above, an open-ended Work Product may either be prompted ("show the mapping across these two representations") or unprompted. When prompted, the student will be asked to construct or complete a second representation given information in terms of a first representation. When not prompted, the Work Product is the trace, the final or intermediate products, and/or a solution protocol from an open-ended solution in which the student may or may not have provided evidence. The directive, "be sure to show your work," helps ensure that the representations will be provided as long as the student is sufficiently familiar with the task format and expectations.

Potential Observations address correctness and quality of the required mappings or explanations of symbolic expressions. Multiple aspects may be evaluated. When the Work Product includes an explanation, Potential Observations include an explanation of what is common across the systems and what differs and how it matters for carrying out what kinds of reasoning.

8.5 Some Connections with Other Design Patterns

Model articulation will often be pertinent in multiple-step tasks, after the model formation step. There are several reasons for this. First, as revealed in Larkin (1983), experts' use of powerful symbol-system representations is generally preceded by the formation of a model in the semantic layer—that is, in terms of the entities, relationships, and processes in the model. These are the connection to the symbol system, rather than the symbol system representation being mapped directly to the situation. Any steps of model-based reasoning carried out with symbol-system representation have necessarily required model articulation to set the stage.

Second, solving a problem often requires transforming information about a real-world situation from expression in one system to another for a different purpose. A table may be a good way to represent the outcome of an experiment, but this representation is not optimal for quantitative manipulations in the way that an algebraic expression or statistical graphic would be. The genetics example shown in Box 4 illustrates this point. The results of crosses are shown as tallies, but they must be transformed to the representation of a Punnett Square in order to bring to bear the machinery associated with this standard form for reasoning about the results of crosses.

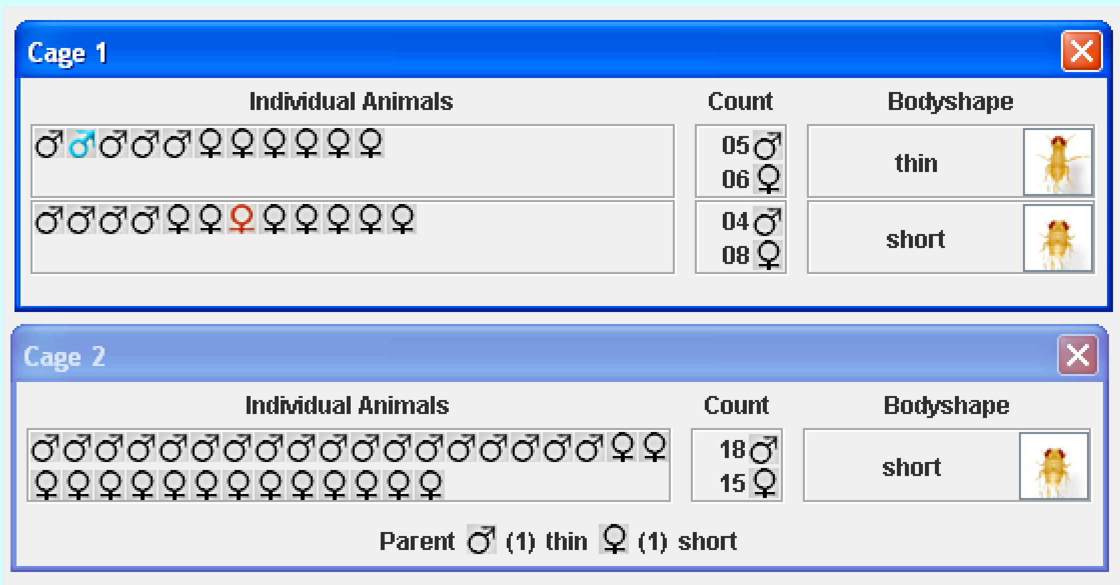
Third, when the results of symbol-system operations are completed or when the outcome of an investigation is summarized, these outcomes must be expressed in a form that communicates the outcomes in term of the model's semantics. Articulation to a representational form that is tuned to communication, which is generally not the same as the form that is tuned to the operations, is required in these cases. Labeled path diagrams are used to report the result of structural equations modeling, for example, while matrix algebra was the representation through which estimation was carried out.

Box 4. Model-Based Reasoning Tasks in Genetics: Model Articulation

An instructor interested in determining how well students understand a given model can use a Model Articulation task to see if they can reason across representations. In this genetics example, students are presented with the Virtual Genetics Lab representation of a cross as shown below. Students may then be presented the following tasks:

1) Use a Punnett Square explain how the results of the cross were obtained.

2) What is the expected percentage of offspring with a short body type? How did you obtain that answer?



For this problem students must be able to articulate how a cross is performed, and must understand the relationship between the results given for cage 2 and the entries of a Punnett Square. They must then be able to move from a graphical representation to a numerical representation in terms of the percentages associated with the possible body types.

As with all model articulation tasks, this type of problem requires that students to be familiar with multiple representations of the subject matter. The focal KSA is the transition between different representations. For this problem three different representations are given; an instructor could remove one or add more to decrease or increase difficulty. The difficulty will be affected by how familiar students are to each of the representations.

Work products would include the Punnett square produced by the student and the student's explanations. From these work products an instructor could determine how well students understand the concept of crossing and how well they are able to use multiple representations to obtain conclusions about the results of the given cross.

9.0 Model Evaluation

9.1 Rationale, Focal KSAs, and Characteristic Task Features

Model evaluation is examining the appropriateness of a model for a situation or a set of data. This may be as straightforward as addressing the binary question regarding whether or not the model fits the data, or it may involve an investigation of how well or in what respects the model fit. While tasks can be devised that focus primarily on model evaluation, this aspect of model-based reasoning is intimately connected with several other aspects of model-based reasoning, and it is always part of a fuller investigation that uses models. Model evaluation is tied inextricably with model use. In order to evaluate a model, students must be able to reason through the model to put forward its facsimile of the salient features of the situation, whether qualitative or quantitative, because comparing these projections with the actual situation is the basis of model evaluation. While it may be hard to separate model use and model evaluation (and often unnecessary, indeed undesirable), tasks can be designed to focus on model evaluation for more targeted instruction and assessment.

In any type of model-based reasoning, students need to be able to make the connection between the real-world situation and the model (this is the arrow in the lower left corner of Figure 1), as discussed in Section 5 on model formation. Without model evaluation, students have no justification for why one model may be better than another, and therefore may not be able to determine an appropriate model. In real-world situations where the model is not provided, students will have difficulty addressing the problem if they cannot evaluate (as well as propose) candidate models.

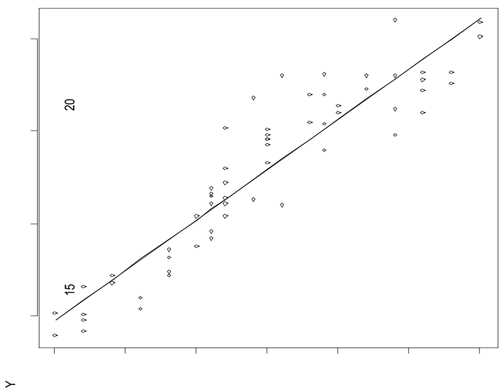
There are three main forms of model evaluation tasks: one in which the model is provided and the student must address its appropriateness; another in which multiple candidates are provided or suggested and the student must determine what model should be used; and finally, tasks in which the model is not given and the student must formulate a model. The first two forms focus attention on model evaluation specifically, while the last of these brings model evaluation into the flow of all or part of the inquiry cycle. In an investigation, a researcher can have a rubric that includes assessing model evaluation as it occurs in students' ongoing procedures or in their final presentations.

Model evaluation is often prerequisite to model revision and model elaboration; it is necessary to determine how and how well a model fits a situation before one can improve it. In physics, the quantum revolution at the beginning of the twentieth century was motivated in part by the failure of Newtonian and field mechanics to account for the photoelectric effect and "block box" radiation. Tasks designed to provide evidence about model evaluation can be naturally extended by following up with model revision or elaboration.

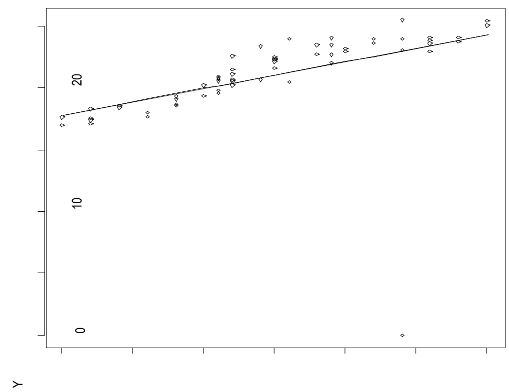
One classic example of model evaluation in statistics is the use of multiple regression. In multiple regression, a model stipulates the relations among a collection of variables in which a set of independent variables is used to predict an outcome or dependent variable. How well the model fits and the structure of the relationship of the dependent variable given the independent variables are studied with a variety of model-checking tools (Belsley, Kuh, & Welch, 1980; Mosteller & Tukey, 1977). .

In the special case of simple regression, the model posits a linear relationship between the two variables, or $y = a\,x + b$. For example, an analyst may hypothesize that age and strength are related, such that as people grow older they get stronger. In order to test this theory, or evaluate the linear relations model, a researcher would obtain data regarding subjects' ages and strength. Once data have been collected, several methods can be used to study the fit of the model. One method is to test the theory graphically (note the articulation between an equation representation and a graphical representation). The researcher can graph the data points and see if the results look like the expected graph of variables with the theorized relationship. In the case of age and strength, the researcher may find that the graph looks more curvilinear (at some point as people age they start to lose strength), and may decide that their model of a linear relationship does not apply or only applies to a given range of ages (thus moving in the inquiry cycle to model elaboration or model revision). The three plots in Figure 14 show examples of data for which the linear regression model appears suitable, a curvilinear relationship that it cannot capture, and a relationship with an outlier that renders the least-squares regression line seriously misleading. The second two plots are examples of qualitative patterns of misfit that statistics students must learn to recognize in order to evaluate the fit of regression models.

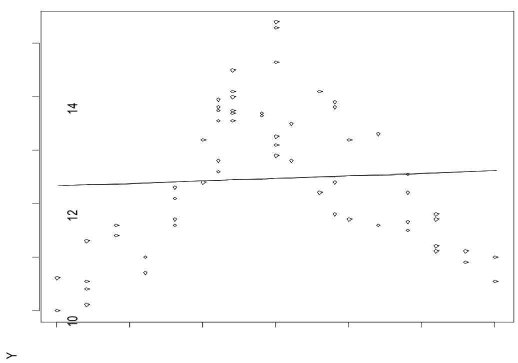
Figure 14: Examples of a Simple Regression Model with Three Data Situations



Regression is suitable



Outlier in regression model

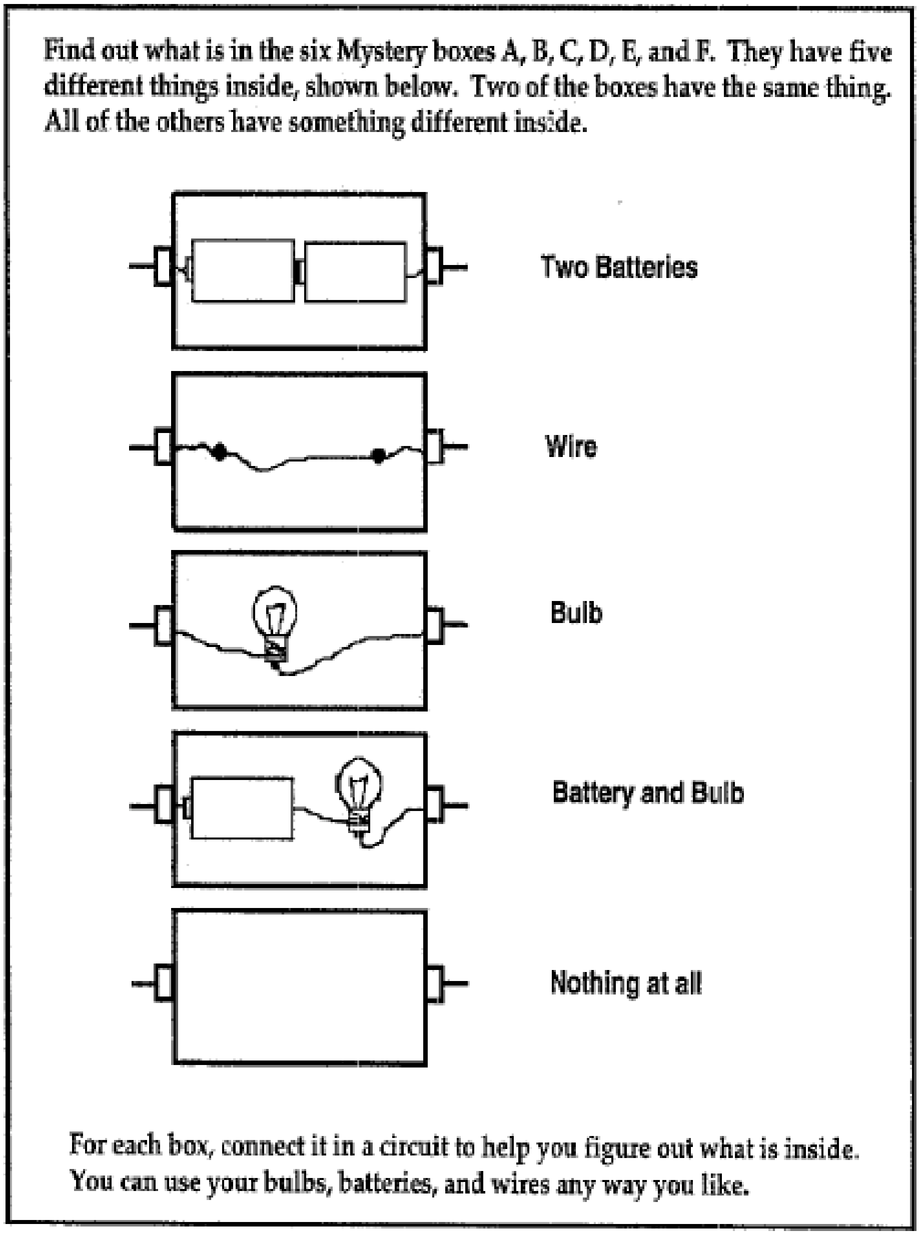


Curvilinear relationship

The researcher can also use statistical methods of testing to examine model fit. An r^2 value can be calculated to quantify the proportion of variance of the dependent variable that is explained by the independent variable(s). This value can be used as one quantitative measure of how well the data fit the model, both in absolute terms as to how well the dependent variable is predicted and in comparative terms as to how much better a curvilinear model fits, for example, than a linear model.

An example that combines model evaluation, model use, and model revision is Baxter, Elder, & Glaser's Mystery Boxes (1996). In this task, students are given six different boxes with some combination of elements among a light bulb, wire, and batteries, and they must perform tests in order to determine what is inside the box (Figure 15).

Figure 15: A Mystery Box Task (Baxter, Elder, & Glaser, 1996)



The students have been studying a model for simple circuits with these kinds of components. In this hands-on task, they must use their understanding of this model to determine what sub-model fits each of the boxes. They must determine which tests (connecting the terminals of the mystery box with just a wire, with a battery, with a light bulb, and so on) are appropriate in narrowing down the choices for the submodel. Interpreting the results of a test requires reasoning through each posited model to predict what would be observed if it were the true configuration (model use), then determining whether the observed result is consistent with the prediction. The comparison between this prediction and what actually happens is an instance of model evaluation. Generally a single test is not sufficient to determine conclusively which configuration is inside a box, so the results of multiple tests must be synthesized to evaluate each possibility. This feature of the task leads to some Potential Observations concerning evaluation strategies that will be discussed in Section 9.4.

The Focal KSAs in model evaluation tasks are the capabilities to determine whether, how well, or in what aspects, a model is appropriate for a given situation — either a real-world scenario or an already-synthesized data set. This can be elaborated to encompass identifying relevant features of the data and the model(s) under investigation and evaluating the degree and/or nature of the correspondence between them. In the tasks discussed above, students must be able to examine the data given (the variables in the regression data set or the mystery box) and determine either whether their model fits or which model is appropriate.

Characteristic Features of every task designed to assess model evaluation include a target situation and one or more models. The models should be able to be examined in light of the situation and the data given. In the regression example, the situation is the being able to predict an outcome variable, and the model is the statistical relationship between the variables given. In Mystery Boxes, the situation is determining what circuit in a given box produces the observations that the set of tests reveals, and the family of models at issue is the set of completed circuits that can be formed from the configurations of elements within the boxes and elements that can be used to connect the terminals.

9.2 Additional KSAs

In addition to the capabilities described above, assessments of model evaluation can require different levels of domain knowledge and familiarity with the type of task or model being evaluated. With regard to domain knowledge, being able to evaluate the fit of the model depends on being able to identify mismatches between a model and a situation. The more subtle the mismatch and the more it depends on the particulars of the model, the more critical the Additional KSA of domain knowledge becomes, because domain knowledge sets expectations about what features are relevant or irrelevant and what relevant patterns should look like. Thus an assessment meant to focus on the process of model evaluation per se rather than model revision as part of learning about particular models ought to use familiar models and situations. An assessment meant to focus on model revision at the same time as knowledge of a particular model can validly have an appropriately high requirement for the substantive issues that are involved.

Different levels of familiarity also will be needed regarding the method used to evaluate the model, as well as with the standards and expectations in the field. As noted below with respect to Variable Task Features, a task designer can employ scaffolding to reduce the demand on Additional KSAs consisting of background knowledge or planning

In the regression task, for example, students may be asked to use different methods for evaluating model fit, such as graphical or statistical, and may be asked to use tools that help carry out evaluation procedures but introduce their own knowledge requirements. For instance, they may be asked to calculate fit indices by hand, which could increase demands for computational procedures, or they may be able to use a computer program, which reduces demand on computation but adds the Additional KSAs required to interact with the program.

In the Mystery Boxes task, students' knowledge of circuits materially affects the difficulty of the task. In the Baxter et al. study, the students had just completed a unit on electrical circuits, so the focus of the task for them was in planning and carrying out the testing procedures to figure what was in each box. Students who are not familiar with circuits might not be able to reason through possible combinations of elements (model use) to carry out model evaluation, although the tasks could then be used as instructional activities to help them develop knowledge about electrical circuits. Students also will be required to have some knowledge of how to hook up the different components; these are Additional KSAs concerning procedures that would be circumvented in a computer-based version of the task. Furthermore, students could be told how to proceed in evaluating the boxes in order to reduce demands on planning and organizational capabilities. Baxter et al. chose not to provide scaffolding because determining whether the students could plan and rationale their testing procedures was of interest in their research.

9.3 Variable Task Features

Model evaluation tasks can vary as to the type and complexity of the model or models to be examined. Some situations may present the models to the student, and the student will be explicitly directed to examine them, while in other situations students may have to determine the model(s) themselves, and indeed whether or not to evaluate fit. That is, model evaluation can be prompted or unprompted (implicit) in a given task. Whether or not the model fits, and the degree and nature of misfit, also can be varied. The type and complexity of the model evaluation methods may differ. Each choice can either highlight or minimize demands for particular aspects of the Focal KSAs and Additional KSAs, as discussed above. Different choices can provide more or less information, often trading off against convenience and economy of scoring procedures. For example, more open-ended tasks can take longer for students to complete and present more challenges for scoring, but provide more information about students' reasoning and incorporate model evaluation into more authentic situations.

In the regression task, for example, the number of variables used for prediction can be varied. The students can also just be asked to use graphical displays to explain why they believe the model fits or does not fit, or they can be asked to use statistical methods or graphical methods to justify their conclusions.

In the Mystery Boxes example, the medium of the task could vary as students could have the physical boxes to work with, an interactive computer simulation of the boxes, or paper and pencil static representations. The last of these is simplest and easiest to score, but it places a greater demand on model use for projecting the results of different configurations because the task environment itself no longer provides feedback. In addition, the number of boxes and the content of the boxes could be modified, as well as the amount of information the students have regarding what might be found in these boxes. All of these modifications can affect the difficulty the tasks and highlight or downplay the importance of different aspects of knowledge, both focal and additional.

The Mystery Boxes assessment illustrates another design choice that is affected through Variable Task Features, namely the kind and amount of scaffolding provided. In order to determine what configuration is inside the box, the student needs to formulate and reason through an electrical circuit model. To focus more directly on model evaluation, the student can be given scaffolding such a chart that lists the results of different tests when applied to different configurations. Note that this scaffolding would remove most of the requirements for reasoning through the electric circuit model per se and shift the focus to strategies and efficiencies in model-evaluation procedures.

9.4 Potential Work Products and Potential Observations

The simplest Work Product for a model evaluation task is the indication (in whatever format specified) of whether or not the model fits or which model fits. It is also least informative. The next more informative option is having the student provide qualitative and/or quantitative indications of degree and nature of fit and misfit. Statistical tests, graphs, or other representational forms for model evaluation can be evoked as Work Products. These “final product” Work Products can be accompanied by an explanation (verbal or written) of why and how the student reached the conclusion. This can include verbal or written explanations of the hypotheses formulated (regarding model fit) and the methods used to test this hypothesis, including the output from formal model fitting tools. Written, verbal, or computer-tracked traces of the actions that the student performed can also be collected as Work Products.

Compared with simple choice Work Products, explanations are particularly useful in determining if a student understands the situations and models well enough to evaluate them critically. The formality of the assessment may dictate the format required as well as the amount of depth expected. When additional tools are required, the explanations also can be useful in assessing the students’ understandings of these tools—Additional KSAs in that a designer can choose to avoid them, but in such cases can choose to incorporate them as targets of inference.

The trace of a solution can take various forms, such as a computer trail of the actions that the student took through a computer interface in a simulation-based investigation, a video recording of the actual performance, or a written trace by the student of the order in which they performed each of the steps of their evaluation. All of these examples provide more evidence about the efficiency of the student’s model evaluation procedures than a final solution. Baxter, Elder, and Glaser (1996) found the last of these particularly useful for evaluating the rationale behind the students’ decisions.

In model evaluation tasks, Potential Observations can thus address the comprehensiveness and the appropriateness of the hypothesis generated through the model for evaluation, the appropriateness of the evaluation method(s) used to assess model fit, the efficiency and the adequacy of procedures the student selects, and the correctness and thoroughness of the evaluation. In particular, the following features of students' work can be evaluated:

- Whether students identify cues of model misfit
- Whether particular areas, patterns, or unaccounted-for features of the situation are identified
- Whether hypotheses for the model-data discrepancy can be proposed

All of these Potential Observations provide evidence about model evaluation in context, but all also require some degree of understanding of the substance of the situation, as noted above in the Additional KSA discussion.

More complex Work Products provide the opportunity to explore these qualities more deeply and allow several aspects to be evaluated. Simpler Work Products—such as selection of a best-fitting model—provide less information but, on the other hand, allow the aspect of proficiency to be targeted more precisely. The quality of the explanations given, as well at the quality of the determination of how an ill-fitting model might affect inferences resulting from that model, also can be examined. How well a student is able to integrate the results from different methods of testing fit can be observed along with how well a student is able to indicate which parts of either the model or the data do or do not fit.

In the regression example, the Work Products can include the output from a formal model fitting tool and an explanation of the conclusions drawn by the student from observing this output. In the simple case, the output may also be a graphical representation of the data. From these outputs, one could observe the quality of the explanation and the appropriateness of the tools used and their application.

In the Mystery Boxes tasks, Baxter et al. gathered as Work Products students' initial plans, strategies, and explanations of their solutions, and traces of their activities in the form of talk-aloud protocols. The researchers evaluated their explanation Work Products for what they believed they would see if a certain combination of materials was inside the box—i.e., the aspect of model use that would be integral to evaluating a proposed model. From the trace of students' activities, the investigators could observe and evaluate how flexible the students were when they monitored the results they were seeing. The Work Products gathered made it possible to create observations that addressed not only the end results (the students belief about the contents of each box) but also how well the students were able to interpret the results of each of the tests in order to determine which or if further tests are needed.

9.5 Some Connections with Other Design Patterns

As mentioned above, the Model Evaluation *design pattern* is tied closely with model use. In some situations it may be difficult to develop tasks that measure only model evaluation; however, tasks can be designed to emphasize either model use or model evaluation or both. This may be accomplished by scaffolding whichever aspects of reasoning (if any) are not the intended focus of the task (e.g., providing a table of test results in the Mystery Box tasks).

The Model Evaluation *design pattern* also is associated with model revision and model elaboration because, generally, in order to determine if a model needs to be revised or elaborated, some model evaluation needs to have been performed. Model revision and model evaluation tasks can be designed to eliminate the step of model evaluation by presenting the students with a situation and a model they are told is inadequate in some particular way, one that they need to revise or elaborate on.

10.0 Model Revision

Model revision is the aspect of model-based reasoning that allows us to speak of the inquiry *cycle* rather than the inquiry *sequence*. We form a model for a situation and purpose, we reason through it to evaluate its aptness—and more often than not, find that in some way it isn't quite right. We must then use the clues about just where and how the model doesn't fit to modify it and improve the correspondence to better serve our purposes. Model revision is closely related to model evaluation because it is necessary that inadequacies of a provisional model be identified.

10.1 Rationale, Focal KSAs, and Characteristic Task Features

The Focal KSA in model revision is the capability, in a given situation, to modify a given model so that its features better match the features of that situation for the purpose at hand. This capability can be further differentiated into recognizing the need to revise a provisional model, modifying it appropriately and efficiently, and justifying the revisions in terms of the inadequacies of the provisional model.

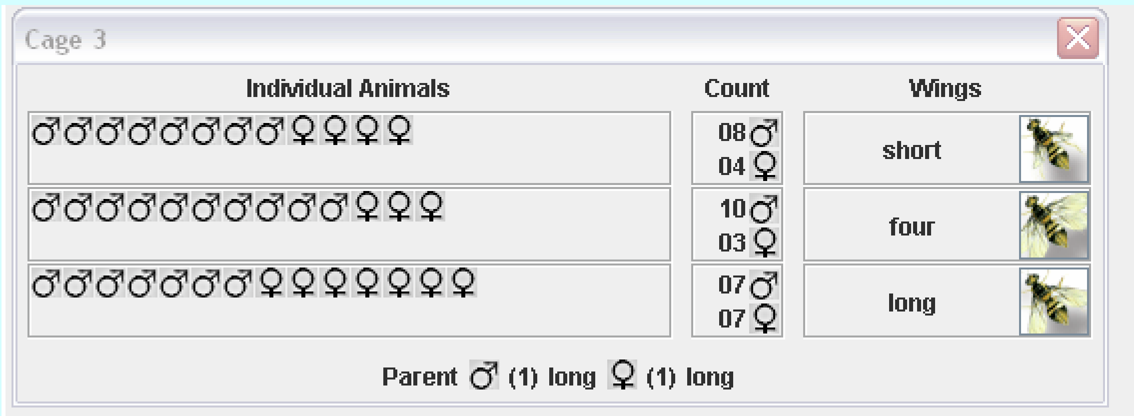
Model revision tasks feature a situation to be modeled, a provisional model that is inadequate in some way, and the opportunity to revise the model in a way that improves the fit. Box 5 presents an example of a task using stimulus material from the Virtual Genetics Lab that requires model revision. The structure used there is to present the examinee with a situation and a model that has been proposed by a hypothetical student. The examinee must evaluate and then revise the provisional model.

Another example of a task that used an adaptive procedure to elicit evidence about model revision was suggested in the Biomass project (Steinberg et al., 2003). A student would be provided the results of a first crossing of animals with an unknown heredity structure for the trait of interest. These results would be constructed to be consistent with two different inheritance structures. The student would be asked to propose one plausible model for the dominance structure among the alleles. The results of a second crossing would contain enough information to distinguish between the models that fit the first crossing — and the second-crossing results presented to each student would be selected to be inconsistent with that student's first response and consistent with the one that she did not propose. Model revision would be required no matter what the student first hypothesized.

Box 5. Model-Based Reasoning Tasks in Genetics: Model Revision

A task schema that can be used to assess Model Evaluation and Model Revision is to present an examinee with an incorrect model proposed by a fictitious student. The ways in which the provisional model is incorrect are chosen to highlight whatever features of the substantive model or the evaluation techniques are the target of inference. The task illustrated here was developed by the authors of this paper, but uses a representation from the Virtual Genetics Lab to illustrate the approach.

The background for this task explains that a student Jose has found six bugs in a shed, four with long wings and two with short wings. He decides to investigate the mode of inheritance of wing type. He hypothesizes that there are two alleles and the mode of inheritance is simple dominance. He crosses two long-winged bugs. To his surprise, he obtains the following offspring:



Model Evaluation is first required. The appearance of a third wing type, four wings, contradicts Joe's hypothesis. Further investigation will show that while there are in fact two alleles, the mode of inheritance is co-dominance: when the two different alleles are combined they produce a third variation of the trait. The data shown above are not conclusive, so repeated cycles involving model formation, crossing, and model evaluation will be required.

10.2 Additional KSAs

As with model evaluation, the design of model revision tasks requires thought about the involved domain knowledge. We saw that especially in advanced and complex tasks, an understanding of the scientific phenomenon at issue is increasingly important in detecting anomalies and inadequacies because it sets up expectations for key entities and relationships. In these situations, the same knowledge is instrumental in revising the model to deal with the inadequacies: What patterns are not being modeled appropriately, and how might they be modeled? The following section on Variable Task Features discusses approaches that a task developer has for addressing the relationship between domain knowledge and model-based reasoning in model revision tasks.

As usual, familiarity with task type, tools and representations, and expectations for performance are Additional KSAs that can make the difference between success and construct-irrelevant failure on a model revision task. Variable Task Features should be controlled in a way that is appropriate to the context and purpose of assessment to remove such alternative explanations for poor performance (e.g., by presenting students tasks whose demands in these respects they can handle).

10.3 Variable Task Features

Model revision tasks can vary as to model and substance—and the students' familiarity with task format, tools, representations, and expectations. The features of model revision tasks can be varied systematically in order to manage demands for the Additional KSA of domain knowledge. Within the heading of "familiarity," we can distinguish the following approaches.

One way to minimize the demand for domain knowledge in large-scale tests is to make the context simple and familiar, as with everyday experience. While removing the sometimes undesired evidentiary confound of domain knowledge from model-based reasoning, this approach has the side effect of also removing the desirable scientific confound between domain understanding and model-based reasoning.

An alternative is to craft a task that is based on a more substantive scientific model with which the students are known to be familiar. The desired connection between domain knowledge and model-based reasoning is now exploited, while the evidentiary focus is on reasoning, conditional on the required domain knowledge. This approach is consistent with the instructional philosophy that holds that to understand a scientific model necessarily includes being able to reason with it. Note that carrying out this assessment approach requires knowing that the student is sufficiently familiar with the model area that is at issue. Also, it is natural to implement this approach when connected with instructional programs or determined locally by teachers who know what students have been studying.

When it is desired in large-scale testing to employ model revision tasks with substantial demands for domain and model knowledge, evidence about the domain knowledge and reasoning are again confounded. To disentangle them, a task can include multiple directives, some of which address domain knowledge and others of which use domain knowledge in the course of model revision.

A set of related Variable Task Features in model revision tasks are whether a provisional model is provided, inadequacies of the provisional model are provided, model revision is prompted, and the task situation is interactive. A task that focuses exclusively on model revision provides a provisional model, points out its inadequacies, directs the student to revise it accordingly, and provides no further iteration. This specificity is sometimes exactly what is desired, perhaps to focus attention on revising a particular model in the course of instruction or to obtain a nugget of evidence about a particular educational objective. However, the specificity trades off against the natural application of model revision in conjunction with model evaluation in particular — and model formation and model use more broadly. A second cycle of model evaluation for the revised model is not observed, for example, despite its pivotal role in the inquiry cycle. At the other extreme is seeking evidence in the course of a broader investigation, with sufficiently rich work products to reveal evidence about model revision if it occurs, and rubrics to evaluate its quality in terms of Observable Variables. Between the extremes are structured tasks, such as those discussed in White & Frederiksen (1998), that support the student working through the phases of an investigation and in this way prompt for model revision.

As with the other model-based reasoning *design patterns*, model revision has as Variable Task Features the substantive content, type, and complexity of the model at issue, and the representations and tools that are involved.

10.4 Potential Work Products and Potential Observations

Work Products for model revision tasks can include the choice or the construction of a representation of the revised model, and an indication of the problem with the initial model and how modifications could address the issue. Explanations of how the model was revised specifically in response to the ways in which it was found inadequate can also be required, again in either choice or constructed formats.

If model revision is not prompted, as in the form of an unstructured investigation, a more comprehensive Work Product is required; for example, a solution trace, intermediate products, or an explanation of steps taken, so that evidence will be available as to whether model revision was carried out — and if so, its mode and results.

To produce values of Observable Variables from performances to specific tasks, these Work Products can be evaluated for the appropriateness of the methods and the modifications to the model. The quality of the basis on which students determine that their new model is an improvement also can be evaluated, particularly focusing on the degree to which the inadequacies of the original model have been addressed. A multiple-choice task to this end could offer possible corrections and reasons of varying qualities, while an open-ended task would require soliciting a student's rationale and then evaluating its quality with a rubric.

10.5 Some Connections with Other Design Patterns

Because model revision is so central to inquiry, it is worth having a Model Revision *design pattern* to specifically focus on it: Under what conditions can we get evidence about students revising models so we can build tasks with these features and so we can recognize those situations within more complicated activities? What are ways we can capture evidence about students' thinking about how and why to modify models, and what aspects of their work should we call out for evaluation? Yet because of its very centrality, model revision is difficult to assess in isolation from other aspects of model-based reasoning. Model revision is prompted only by model evaluation, as we must first decide that a provisional model is in some way inadequate. We must then use model formation to propose alternatives or modifications that may better address the situation at hand. We must use the revised model to reason forward to its implications for observations that we hope will be in better accord with the situation, and use model evaluation again to determine whether this is so. Perhaps better than any other aspect of model-based reasoning, the Model Revision *design pattern* calls to our attention that these *design pattern*s correspond to distinguishable components of tasks rather than distinct psychological abilities.

11.0 Model–Based Inquiry

Distinguishing aspects of reasoning is useful in instruction and assessment, but it is their coordinated use that marks model-based reasoning in practice. We would like to help students move back and forth among these aspects of reasoning, often without clear demarcation, to understand systems and act effectively through models of them. The general *design pattern* for model based inquiry subsumes the *design patterns* for each of the aspects and calls attention to the coordination among them. More than any of the individual aspects, model-based inquiry highlights the importance of metacognition in moving effectively through cycles of inquiry. This section draws on White and Frederiksen (1998) and White, Shimoda, and Frederiksen (1999).

11.1 Rationale, Focal KSAs, and Characteristic Task Features

The philosophy of science, Giere (1994) argues, assumes that the language of science has a syntax, a semantics, and, finally, a pragmatics. He continues,

> While syntax is deemed important, semantics, which includes the basic notions of reference and truth, has received the most attention. Much of the debate regarding scientific realism, for example, has been conducted in terms of the reference of theoretical terms and the truth of theoretical hypotheses. Pragmatics has been largely a catchall for whatever is left over, but seldom systematically investigated. I now think that this way of conceiving representation in science has things upside down (p.742).

Model-based reasoning, as described by researchers such as Stewart and Gobert, is all about pragmatics. A philosophy of science is not sufficient for either understanding how scientists use models in practice nor for how to help students learn to use them; a cognitive psychology of science is required as well. While the preceding sections on aspects of model-based reasoning have illuminated important cognitive activities in model-based scientific inquiry, it is the heuristics, the strategies, the procedures, and the self-regulating tools that people need to use models effectively in real-world situations. It is this higher-level, coordinating, or executive level of cognition that the Model-Based Inquiry *design pattern* addresses.

The Focal KSAs in this *design pattern* are students' capabilities to manage their reasoning in inquiry cycles. The specific, more technical, aspects of model-based reasoning discussed in the preceding sections are brought to bear, but is their use coordinated, efficient, coherent, and effective—or is movement through the investigation disjointed, unsystematic, inefficient, and aimless? Are students bringing to bear self-monitoring skills to understand whether model evaluation is needed, or does a provisional model need to be revised or elaborated?

Any task developed for an overall assessment of model-based reasoning must contain more than one feature from the more specific *design patterns*. As with all of these *design patterns*, there must be a real-world problem being addressed. This problem must require the use of models and/or a modification of models in order to develop an explanation or prediction of some phenomena. The way that the Model-Based Inquiry *design pattern* goes beyond the specific *design patterns* is that it concerns information and reasoning across the more specific aspects.

Many of the examples mentioned in the previous sections can be expanded to include multiple aspects of model-based reasoning and would therefore apply to the overall *design pattern*. Stewart and Hafner's genetics curriculum can be thought of as one large assessment task, or it can be broken down (as we have seen) into several distinct assessments. In this case, the assessment would start out where the students are applying the simple dominance model to a given situation (as seen in model use). The students then are presented with a situation where this does not fit — as in they are given three possible traits instead of two. The students must then identify the inadequacies of the simple dominance model (model evaluation) and modify their model in order to fit this situation (model elaboration.) Students next are given more information that will lead to more complicated models. At some points they have to elaborate their model further or revise their model as described above as new data are obtained. The Work Products for this overarching task would include the explanations for the models and how they fit the situations, the overall outcomes of using the model to explain or predict different behavior, as well as the representation of the models themselves. These Work Products can be used to evaluate a student's use of model-based reasoning in the context of modes of inheritance.

11.2 Additional KSAs

As with the other *design patterns*, the Additional KSAs in the *design pattern* for assessing model-based inquiry will include knowledge of the models, context, and scientific content of the task at issue. The mix of these Additional KSAs, if any, that is jointly a target of inference with inquiry itself must be determined in light of the purpose of the assessment and intended test population. Those Additional KSAs that are not part of the target of the assessment should be avoided or supported, or the assessor should ascertain that the tested students are sufficiently familiar with them so that they are not significant sources of difficulty.

11.3 Variable Task Features

Because inquiry tasks encompass the aspects of model-based reasoning addressed in the preceding sections, all of the Variable Task Features for relevant aspects are open for consideration in the larger task. This includes the identification and complexity of the model and which tools and representational forms are used. Some design choices can cut across aspects of the larger task (such as the models and content area that are involved) while others (such as scaffolding) can differ from one aspect to another (e.g., a checklist provided just for model evaluation). Time frame is an important Variable Feature for investigations. Non-trivial investigations can easily take an hour or more, and learning tasks in the classroom can extend to days or weeks.

Choices regarding the content area will be shaped by the purpose of the task that the developer has in mind. In the classroom or as part of a curriculum, for example, the content is likely to be based on the models that are the target of instruction, so the task can hold fairly high demands for this knowledge. This was the case, for example, in Baxter, Elder, and Glaser's (1996) Mystery Boxes study, where the students had just completed a unit in electrical circuits. As part of a high-stakes accountability test in which both the models and the inquiry process are addressed in the standards, demands for both may be imposed, and the Additional KSAs regarding the model and scientific content will jointly be construct relevant. In a large-scale task that is meant to focus on the inquiry process and not be confounded with content, the models and content may be sufficiently familiar to students as to introduce minimal possibilities for poor performance for these reasons; for example, models from middle school standards could be used in a high-school task in order to focus the evidentiary value of the task on the inquiry process.

A particularly important Variable Task Feature to consider in designing inquiry tasks is the degree of scaffolding to provide students as they move from one aspect of an inquiry to another, managing information, evaluating progress, and deciding what to do next. This thinking is central to inquiry and one of the hardest aspects of inquiry for students to learn (and for educators to assess). Research that has been carried out in scaffolding students' *learning* about inquiry holds insights for task designers as well.

White and Frederiksen (1998) describe a sequence of seven instructional tasks that constitute a middle-school course on mechanics in the context of their ThinkerTools software. The amount of scaffolding was progressively decreased as students became familiar with inquiry processes and expectations. In an inquiry assessment task, providing more scaffolding is appropriate for earlier learners; it helps them engage meaningfully with the task and makes sure that some evidence will be obtained for different aspects of the investigation. On the other hand, the more the processes are scaffolded, the less evidence is available about the students' capability to manage their progress through the investigation.

> Associated with each Task Context is a Task Document in which users do their work for that task. For example, there is a Project Journal, a Project Report, and a Project Evaluation, as well as a System Modification Journal (in which users record a history of their system modifications and the reasons for them). These documents are organized around a possible sequence of subtasks (or subgoals) for that task. For example, the Project Journal is organized around the Inquiry Cycle that we employ in our ThinkerTools curriculum (White, Shimoda, & Frederiksen, 1999, p. 163).

White, Shimoda, & Frederiksen's (1999) computer-based environment SCI-WISE for carrying out investigations also provides interactive support in the form of personified "agents":

> In addition to Task Documents, each Task Context has a set of advisors associated with it, including a Head Advisor and a set of Task Specialists. There is a Head Advisor for each Task

> Context; namely, the Inquirer for doing research projects, the Presenter for creating presentations, the Assessor for evaluating projects, and the Modifier for making changes to the SCI-WISE system. The Head Advisor gives advice regarding how to manage its associated task, suggests possible goal structures for that task, and puts together an appropriate team of advisors. For example, our version of the Inquirer follows the Inquiry Cycle shown in [Figure 2 of this paper]. It suggests pursuing a sequence of subgoals, and each such subgoal has a Task Specialist associated with it, namely, a Questioner, Hypothesizer, Investigator, Analyzer, Modeler, and Evaluator (p. 164).

The advisors can offer assistance both for processes, such as moving effectively through the inquiry cycle, and for products, such as evaluating whether a hypothesis the student has proposed can be tested. For example, Helena Hypothesizer tells the student "Here are some things I can do for you: (1) I can describe the characteristics of good hypotheses; (2) I can suggest strategies for creating hypotheses and advisors who can assist; and (3) I can help you evaluate your hypotheses to see if they need revision" (White, Shimoda, & Frederiksen, 1999, p. 167).

In a computer-based assessment task, a task developer could choose which agents to make available to examinees and what degree of support they could provide, in order to tailor scaffolding both within and between aspects of model-based reasoning during an inquiry task. As always, however, providing tools that support inquiry-related KSAs introduces at the same time a demand for the Additional KSAs to use them effectively. White, Shimoda, and Frederiksen (1999, p. 177) caution that "it is relatively easy to create versions of SCI-WISE that many would find annoying and confusing—annoying in that the system provides too much advice and structure, and confusing in that there are too many agents who are indistinguishable from one another."

11.4 Potential Work Products and Potential Observations

Model-based inquiry tasks can be designed to produce Work Products that provide evidence specific to aspects of model-based reasoning within the investigation and/or evidence about managing model-based reasoning across aspects in the course of the larger investigation. Because Work Products and Potential Observations were discussed in each of the preceding sections, after a brief comment, this section will focus on Work Products and Potential Observations that address the larger inquiry context.

As mentioned above, all of the Work Products discussed in the preceding sections that contain evidence about aspects of model-based reasoning can be considered in a fuller inquiry task, and all of the Potential Observations that could be evaluated for these aspects can be included as well. In a more detailed scoring scheme, the Observable Variables from the specific aspects can be evaluated and reported separately. This practice is particularly useful for providing feedback to students in instructional tasks in order to help shape their learning: what did they do well in this task, where did they have trouble, and what experiences will help them improve?

Work Products that directly evidence the larger inquiry process must be capable of providing information beyond specific aspects of model-based reasoning. This means providing evidence about the way a student moves through the investigation. One class of Work Products provides some kind of trace of the steps a student has taken, such as a video recording, a talk-aloud protocol, or at some level of detail, the actions captured in a computer–based task. The National Board of Medical Examiners' Primum® computer-based diagnostic tests that are now part of the medical licensure requirements in the United States, capture each step in a solution in what they call a "transaction list." Automated scoring algorithms (we'll say more about this below) are used to extract information from the transaction list both about the final solution and selected aspects of the solution process. Less comprehensive Work Products include notebooks, explicit reports of inquiry phases, and written or oral explanations along the way of why certain actions were taken. Oral explanations can be prompted or unprompted as to their content. We will say more below about responses to "metacognitive" kinds of questions.

Work Products that can provide indirect evidence about inquiry procedures in an inquiry task are final and intermediate products. A correct solution presumably is more likely to have occurred from effective model-based reasoning, although the efficiency of that reasoning is not available to evaluate from this Work Product alone. The qualities of a final solution to a problem, such as a model proposed for a situation after multiple iterations through the inquiry cycle, can be of interest in and of themselves. Only qualities of the final product may be addressed when the purpose of an assessment is licensure, for example. But when the purpose is learning, the evaluation of successive provisional models offers clues about the efficiency and appropriateness of successive cycles of model evaluation and revision.

We note that the choice of Work Products to present students with is linked to the choice of scaffolding to provide. The task documents that White, Shimoda, and Frederiksen (1999) provide as the vehicle for students to record, evaluate, and explain their progress through an investigation not only serve as a Work Product, but they support metacognition for managing activity through the investigation.

What Observable Variables can be evaluated from Work Products that hold evidence about model-based inquiry? Baxter, Elder, and Glaser's (1996) Mystery Boxes tasks were used in a research study of "expertise" in middle school students' inquiry capabilities in a domain known to be familiar to them. Table 2 summarizes the differences they found in Work Products in the form of talk-aloud protocols and solution traces. They are the basis of generic Observable Variables that can be applied broadly to inquiry tasks, as tailored to the processes in the specific investigation.

Baxter, Elder, and Glaser (1996) evaluated students' investigation procedures by painstakingly parsing "thick" Work Products such as explanations, solution paths, and conversations of thirty-one students. In more complex investigations at any scale, the amount of rater time and expertise required to carry out these evaluations for these Observable Variables renders them impractical. An alternative that is available when the investigations are carried out in a computer-based form is automated scoring of solution traces (Williamson, Mislevy, & Bejar, 2006). In the National Board of Medical Examiners' Primum® tasks mentioned above, low-level features of solutions are identified, combined into higher-level features through logical rules (such as whether efforts to stabilize an emergency patient were carried out first rather than later in the investigation), and evaluated by means of a regression function that compares them to the high-level features of experts' solutions (Margolis & Clauser, 2006). Similarly, Stevens (1996) uses a neural net approach to evaluate the efficiency of students' investigations in an epidemiology investigation.

Table 4: Quality of Cognitive Activity in Mystery Box Solutions (Baxter, Elder, & Glaser, 1996)

	Range of Variation	
Cognitive Activity	Low	High
Explanation	Single statement of fact or descriptions of superficial features	Principled, coherent
Plan	Single hypothesis	Procedures and outcomes
Strategy	Trial and error	Efficient, informative, goal-oriented
Monitoring	Minimal and sporadic	Frequent and flexible

A class of paired Potential Work Products and Potential Observables that is particularly well-suited to instructional tasks is based on responses to metacognitive questions. That is, these are the questions that students should be learning to ask themselves as they develop their capabilities in inquiry. For earlier learners, the answers to these questions provide evidence about the degree to which they are thinking well about the appropriate features of their work as it proceeds. Perhaps more importantly, their very presence helps the students learn that these are in fact the kinds of questions that are important in inquiry, and they should come to internalize these questions as they have experience with them. For example, White and Frederiksen (1998) acquaint students with a concept they call "Being Systematic," and define it as follows: "Students are careful, organized, and logical in planning and carrying out their work. When problems come up, they are thoughtful in examining their progress and deciding whether to alter their approach or strategy." As a Work Product, students are directed to rate their own solutions with respect to how systematic they were, on a 1-to-5 scale from "not adequate" to "exceptional."

11.5 Some Connections with Other Design Patterns

Model-based inquiry is a larger activity that draws repeatedly and cyclically on the more specific aspects of model-based reasoning. When designing an inquiry task, a test developer can use this *design pattern* to think about the characteristics of Task Features and Work Products that provide evidence about the movement in the larger space, and the specific *design patterns* to ensure that evidence is elicited about the finer details of the investigation as needed.

The iterative testing and repairing that characterizes troubleshooting can be viewed as a special kind of model-based inquiry. Steinberg and Gitomer's (1996) troubleshooting tasks in the hydraulic system of the F-15 aircraft, for example, can be viewed as iterative cycles of model use, model evaluation, and model revision, with the efficiency of diagnostic tests at the crux of evaluation. The efficiency of tests for evaluating a model becomes particularly important in these more complex tasks. Efficiency is intimately related to understanding both the system in question and the tests that can be carried out, both Additional KSAs that are required jointly for effective troubleshooting. Frezzo, Behrens, and Mislevy (2009) showed how *design patterns* for creating troubleshooting tasks in network engineering are used in the Cisco Networking academy. Siebert et al. (2006) presented a more general *design pattern* that encompasses troubleshooting, called "Hypothetico-Deductive Problem Solving in a Finite Space."

12.0 Conclusion

Model-based reasoning, and inquiry in general, are both increasingly important and difficult to assess (Means & Haertel, 2002). Assessing factual knowledge and isolated procedures is easier and more familiar — and not surprisingly, constitutes the bulk of science assessment carried out today. The *design patterns* developed in this presentation can be used as starting points for building assessment tasks that engage more deeply with model-based reasoning. Task developers can determine which aspects of model-based reasoning they want to address and use the corresponding *design patterns* to make them aware of design choices, then support their thinking about how to make those choices. The *design patterns* are organized around elements of an assessment argument structure as it has emerged from recent research on assessment design and validity theory. In this way, the *design patterns* leverage both research on model-based reasoning and practical experience in assessment design in this area, in a form that is specifically aimed to support task developers.

12.1 Standards-Based Assessment

As part of the standards-based reform movement over the last two decades, states and national organizations have developed content standards outlining what students should know and be able to do in core subjects, including science (e.g., NRC, 1996). These efforts are an important step toward furthering professional consensus about the kinds of knowledge and skills that are important for students to learn at various stages of their education. They are the basis of states' large-scale accountability tests, in accordance with the requirements of the 2001 No Child Left Behind (Public Law 107-110, 2002) legislation.

But standards in their current form are not specifically geared toward guiding assessment design. A single standard for science inquiry will often encompass a broad domain of knowledge and skill, such as "develop descriptions, explanations, predictions, and models using evidence" (NRC, 1996, p. 145) or "communicate and defend a scientific argument" (p. 176). They usually stop short of laying out the interconnected elements that one must think through to develop a coherent assessment: the specific competencies that one is interested in assessing, what one would want to see students doing to provide evidence that they had attained those competencies, and the kinds of assessment situations that would elicit those kinds of evidence.

Design patterns bridge knowledge about aspects of science inquiry that one would want to assess and the structures of a coherent assessment argument, in a format that guides task creation and assessment implementation. The focus at the *design pattern* level is on the substance of the assessment argument rather than on the technical details of operational elements and delivery systems. Thinking through the substance of assessment arguments for capabilities such as model-based reasoning and inquiry promotes the goals of efficiency and validity. It enables test developers to go beyond thinking about individual assessment tasks and to instead see instances of prototypical ways of getting evidence about the acquisition of various aspects of students' capabilities.

In the present case, *design patterns* for assessing model-based reasoning have the advantage of building on research in cognitive psychology, science education, and the philosophy of science. Key insights from these diverse areas of study are brought together in a form where they can be brought to bear in designing assessment tasks for both classroom and large-scale assessments. It is a particular advantage of *design patterns* that they are centered on aspects of scientific capabilities, as opposed to task formats or assessment purposes. In this way, the essence of the capabilities in question and building assessment arguments around them is seen as common, with options for tailoring the details of stimulus situations and Work Products to suit the particulars of a given assessment application.

12.2 Classroom Assessment

Design patterns built around national or state science standards constitute a stationary point to connect both classroom and large-scale assessment with developments in science education and the psychology of learning. There is often a disjuncture between classroom assessment and large-scale assessment; *design patterns* help make it clear that it is the same capabilities being addressed in both, although the assessments reflect different design choices about such features as time, interactivity, and Work Products in order to accommodate the different purposes and constraints of large-scale and instructional tests.

The research on model-based reasoning in science education and cognitive psychology comports with a socio-cognitive perspective on the nature of learning and knowledge. Truly "knowing" models in science is not merely echoing concepts and applying procedures in isolation, but using models to do things in the real world: reasoning about situations through models; selecting, building and critiquing models; working with others and with tools in ways that revolve around the ideas in the models. The only way that students develop these capabilities is by using them, first in supported activities that make explicit the concepts, the processes, and the metacognitive skills for using them. Importantly, the examples of assessment tasks we have used to illustrate science *assessment* are drawn from projects whose focus is science *learning* (e.g., White & Frederiksen, 1998; Johnson & Stewart, 2002; Redish, 2003). We hope that these *design patterns* for assessing model-based reasoning help make these advances more accessible to classroom teachers and curriculum developers as well as to researchers and assessment professionals.

12.3 Large-Scale Accountability Testing

The changing landscape of large-scale accountability assessments places extraordinary demands on state and local education agencies. No Child Left Behind legislation requires large-scale testing at the level of the state, with attendant needs for efficient administration, scoring, linking of forms, and cost-effective development of assessment tasks at unprecedented scales. Tasks must address states' content standards. At the same time, educators want tasks that assess higher-level skills and are consonant with both instructional practice and learning science.

It is widely accepted that more complex, multi-part assessment tasks are better suited to measuring higher-level skills. But cost considerations and incompatibility with conventional test development and implementation practices stand in the way of large-scale use. Many states and their contractors have turned to computer-supported assessment task development and delivery to help them meet these challenges. For large-scale assessments, technology-based tasks such as simulations and investigations that best address higher-level skills and support learning have proved difficult and costly to develop, especially when employing procedures that evolved from conventional multiple-choice item development practices (Riconscente, Mislevy & Hamel, 2005).

Traditionally, items for large-scale assessments are developed by item writers who craft each item individually. These items, which have been written by different item writers, are gathered into a pool of items that may potentially be placed on a test form. Typically as many as half of the items do not survive review. This low survival rate is tolerable because of the low cost of developing individual multiple-choice items. It is not economical when applied to the development of more complex assessment tasks of the sort needed to address higher-level skills. Moreover, the thought and problem solving invested in developing any particular item is tacit in conventional item development procedures. The thinking invested, the design challenges met, and the solutions reached remain undocumented and inaccessible to help another item writer — or even the same item writer — developing additional items. This process is untenable in the long run for technology-based tasks that require an order of magnitude more of time and resources, including complex design rationales, than that required in the design of standard multiple-choice items.

Design patterns are an important part of the solution. A completed *design pattern* specifies a design space of elements to assemble into an assessment argument: Focal Knowledge and Skills, Characteristic and Variable Task Features, Potential Work Products, stimulus situations, and evaluation schemes. This design space focuses on the science being assessed and guides the design of tasks with different forms and modes for different situations. *Design patterns*, in turn, ground *templates* for authoring more specific families of tasks.

In the context of large scale accountability assessments, *design patterns* fill a crucial gap between broad content standards and particular assessments tasks, in a way that is more generative than test specifications and which addresses alignment through construction rather than retrospective sorting. The time and analysis invested in creating *design patterns* eliminates duplicative efforts of re-addressing the same issues task by task, program by program. *Design patterns* can be developed collaboratively and shared across testing programs. Each program can construct tasks which, by virtue of pattern, address key targets in valid ways, but make design choices that suit the needs of their specific constraints and purposes. Thus, *design patterns* add value not just for local development but for accumulating experience and debating standards in the state, national, and international arenas.

12.4 Closing Comments

Model-based reasoning is central to science. Research from a sociocognitive perspective on the nature of model-based reasoning and how people become proficient at using it is beginning to revolutionize science education. Assessment is integral to learning, not just for guiding learning but for communicating to students and educators alike just what capabilities are important to develop, and how to know them when we see them. But the interactive, complex, and often technology-based tasks that are needed to assess model-based reasoning in its fullest forms are difficult to develop. The suite of *design patterns* to support the creation of tasks to assess model-based reasoning promise to bring assessment into line with where science assessment needs to be.

References

Allen, D., Kling, G., & van der Pluijm, B. (2005). Global change curriculum, Unit 1a: Introduction to systems dynamic modeling with STELLA. Ann Arbor, MI: Global change Program, University of Michigan. Retrieved August 2, 2007 from //www.globalchange.umich.edu/globalchange1/current/labs/Lab2/Intro_Stella.htm

Azevedo, R., & Cromley, J.G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*, 523-535.

Baxter, G., Elder, A., Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist*, *31*(2), 133-140.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17,* 205-226.

Baxter, G., & Mislevy, R. (2005). *The case for an integrated design framework for assessing science inquiry (PADI Technical Report 5).* Menlo Park, CA: SRI International.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2009). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *The Journal of Science Education and Technology.* Springer Open Access http://www.springerlink.com/content/566p6g4307405346/

Belsley, D. A., Kuh, E., & Welch, R. E. (1980). *Regression diagnostics: Identifying influential data and source of collinearity.* John Wiley, New York.

Cartier, J. (2000). Assessment of explanatory models in genetics: Insights into students' conceptions of scientific models. (*Research Report No. 98-1).* Madison, WI: National Center for Improving Student Learning and Achievement in Mathematics and Science.

Chi, M. T. H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14,* 161-199.

Chi, M. T. H, Feltovich, P. J. and Glaser, R. (1979). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Clement, J. (1989) Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.). *Handbook of Creativity: Assessment, Theory and Research* (pp. 341-381). New York: Plenum Press.

Clement, J. (2000). Model based learning as a key research area for science education, *International Journal of Science Education, 22,* 1041-1053.

Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist, 28,* 25-42.

Cromley, J.G. & Mislevy, R.J. (2004). Task templates based on misconception research. (*CSE Technical Report 646).* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: Grosset/Putnam.

diSessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 15-33). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

diSessa, A. A. (1993). Towards an epistemology of physics. *Cognition and Instruction, 10,* 105-225.

Duncan, R. G. (2006). The role of domain-specific knowledge in promoting generative reasoning in genetics. *Proceedings of the 7th international conference on learning sciences* (pp. 147–153). Bloomington, Indiana: International Society of the Learning Sciences.

Ericsson, K. A . & Kintsch, W. (1995). Long-term working memory. *Pyschological Review, 102*(2), 211-245.

Fauconnier, G., & Turner, M. (2002). *The way we think.* New York: Basic Books.

Frederiksen, J. R., & White, B.Y. (1988). Implicit testing within an intelligent tutoring system. *Machine-Mediated Learning, 2,* 351-372.

Frezzo, D.C., Behrens, J.T., & Mislevy, R.J. (2009). Design patterns for learning and assessment: facilitating the introduction of a complex simulation-based learning environment into a community of instructors. The Journal of Science Education and Technology. Retrieved January 27,2010, from http://www.springerlink.com/content/566p6g4307405346/

Frigg, R. & Hartmann, S. (2006). Models in science. *The Stanford Encyclopedia of Philosophy (Spring 2006 Edition).* Edward N. Zalta (Ed.). Retrieved from http://plato.stanford.edu/archives/spr2006/entries/models-science/.

Gentner, D., & Stevens, A.L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Erlbaum.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*, 742–752.

Gobert, J. & Buckley, B. (2000). Special issue editorial: Introduction to model-based teaching and learning. *International Journal of Science Education, 22,* 891-894.

Gobert, J. (2000). A typology of models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education, 22,* 937-977.

Greeno, J. G. (1989). Situations, mental models, and generative knowledge. In D. Klahr, & K. Kotovsky (Eds.), *Complex information processing*, (pp. 285-318). Hillsdale, NJ: Lawrence Erlbaum.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics, 33,* 107-133.

Heller, P. & Heller, K. (2001). *Cooperative group problem solving in physics.* Pacific Grove, CA: Thomson Brooks/Cole.

Hestenes, D. (1987). Toward a modeling theory of physics instruction. *American Journal of Physics, 55,* 440-454.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30,* 141-151.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51-74). Cambridge, MA: MIT Press.

Ingham, A. M. & Gilbert, J. K. (1991). The use of analogue models by students of chemistry at higher education level. *International Journal of Science Education, 13,*193-202.

IMS Global Learning Consortium (2000). *IMS question and test interoperability specification: A review* (White Paper IMSWP-1 Version A). Burlington, MA: Author.

Johnson, S. K., & Stewart, J. (2002). Revising and assessing explanatory models in a high school genetics class: A comparison of unsuccessful and successful performance. *Science Education*, *86*, 463-480.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Jungck, J. R., & Calley, J. (1985). Strategic simulations and post-Socratic pedagogy: Constructing computer software to develop long-term inference through experimental inquiry. *American Biology Teacher*, *47*, 11-15.

Kalyuga, S. (2006). Rapid cognitive assessment of learners' knowledge structures. *Learning and Instruction, 16,* 1-11.

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

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95,* 163-182.

Kintsch, W (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294-303.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* New York: Cambridge University Press.

Kintsch,W.,& Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review, 92,* 109–129.

Larkin, J. (1983). The role of problem representation in physics. In D. Gentner & Stevens, A. L. (Eds.). *Mental models* (pp. 75-98). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lehrer, R., & Schauble, L. (2006). Cultivating model-based reasoning in science education. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 371-388). New York: Cambridge University Press.

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D.M. Williamson, R.J. Mislevy, and I.I. Bejar (Eds.), *Automated scoring of complex tasks in computer based testing* (pp. 132-167). Mahwah, NJ: Erlbaum.

Marshall, S. P. (1993). Assessing schema knowledge. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 155-180). Hillsdale, New Jersey: Lawrence Erlbaum.

Marshall, S. (1995). *Schemas in problem solving.* Cambridge University Press, New York.

Means, B. & Haertel, G. (2002). Technology supports for assessing science inquiry. In *Technology and Assessment: Thinking ahead* (pp. 12-25). National Research Council. Washington, DC: National Academy Press.

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

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk, 2,* 237-258.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 257-305). Westport,. CT: American Council on Education/Praeger.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83-108). Charlotte, NC: Information Age Publishing.

Mislevy, R., Hamel, L., Fried, R., G., Gaffney, T., Haertel, G., Hafter et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.

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

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Johnson,. L., & Almond, R. A. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15, 363-378.*

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics.* Reading, MA: Addison-Wesley.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council (2000). *How people learn: Brain, mind, experience and school* (2nd ed.). Committee on Developments in the Science of Learning. J.D. Bransford, A.L. Brown, & R. Cocking (Eds.). Washington, DC: National Academy Press.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser, & N. Chudowsky (Eds.). Washington DC: National Academy Press.

Norman, D. A. (1993). *Things that make us smart.* Boston: Addison-Wesley.

Novick, S., & Nussbaum, J. (1981). Pupils' understanding of the particulate nature of matter: A cross-age study. *Science Education, 65*, 187-196.

Redish, E. F. (2003) *Teaching physics with the physics suite.* Hoboken, NJ: Wiley.

Richmond, B. (2005). *An introduction to systems thinking, featuring STELLA.* isee systems: Lebanon, NH.

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153-196). New York: Academic Press.

Rumelhart D. E., & Norman D. A. (1977). Accretion, tuning and restructuring: Three modes of learning. In Cotton J. W. & Klatzky R. L. (Eds.), *Semantic factors in cognition* (pp. 37-54). Hillsdale, NJ: Erlbaum.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in E-Learning: A framework for constructing “Intermediate Constraint” questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4*(6). Retrieved July 17,2009, from http://www.jtla.org.

Seibert, G., Hamel, L., Haynie, K., Mislevy, R., & Bao, H. (2006). Mystery powders: An application of the PADI Design System using the Four-Process Delivery System *(PADI Technical Report 15*). Menlo Park, CA: SRI International.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24,* 223-258.

Steinberg, L. S., Mislevy, R. J., Almond, R. G., Baird, A. B., Cahallan, C., Dibello, L. V. et al. (2003). *Introduction to the Biomass Project: An illustration of evidence-centered assessment design and delivery capability. (CSE Technical Report 609)*. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Stevens, R. H., Lopo, A. C., & Wang, P. (1996). Artificial neural networks can distinguish novice and expert strategies during complex problem solving. *Journal of the American Medical Informatics Association, 3,* 131-138.

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

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.), *Handbook of research on science teaching and learning* (pp 284-300). New York: MacMillan.

Stewart, J., Hafner, R., Johnson, S., & Finkel, E. (1992). Science as model-building: Computers and high school genetics. *Educational Psychologist*, *27*(3), 317-336.

Stewart, J., Passmore, C., Cartier, J., Rudolph, J., & Donovan, S. (2005). Modeling for understanding science education. In Romberg, T., Carpenter, T., & Dremock, F. (Eds.). *Understanding mathematics and science matters* (159-184). Mahwah, N.J.: Lawrence Erlbaum.

Suárez, M. (2004). An inferential conception of scientific representation. *Philosophy of Science, 71,* 767-779.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87,* 449-508.

Thornton, R. K., & Sokoloff, D. R. (1998) Assessing student learning of Newton’s laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66,* 338-351.

Martin, J. D., & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141-165). Hillsdale, NJ: Erlbaum.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.

Velleman, P. F. (1997). *DataDesk Version 6.0 Statistics Guide.* Ithaca, NY: Data Description.

Wertsch, J. V. (1998). *Mind as action.* Oxford: Oxford University Press.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3-118.

White, B. Y., Shimoda, T. A., & Frederiksen, J. R. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education, 10,* 151-182.

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

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.) (2006). *Automated Scoring of complex tasks in computer-based testing.* Mahwah, NJ: Erlbaum Associates.

Wynne, C., Stewart, J., & Passmore, C. (2001). High school students' use of meiosis when solving genetics problems. *The International Journal of Science Education, 23*(5), 501-515.

Appendix: Summary Form of Design Patterns for Model-Based Reasoning

	Model Formation	**Model Use**	**Model Elaboration**	**Model Articulation**
Summary	This design pattern supports developing tasks in which students create a model of some real-world phenomenon or abstracted structure, in terms of entities, structures, relationships, processes, and behaviors.	This design pattern supports developing tasks that require students to reason through the structures, relationships, and processes of a given model.	This design pattern supports developing tasks in which students elaborate given scientific models by combining, extending, adding detail to a model, and/or establishing correspondences across overlapping models.	Tasks supported by this design pattern assess students' ability to articulate the meaning of physical or abstract systems across multiple representations. Representations may take qualitative or quantitative forms. This DP is relevant in models with quantitative and symbolic components (e.g., connections between conceptual and mathematical aspects of physics models)
Rationale	Ability to build a model is a fundamental component of inquiry-based science. During the construction of a model, students make design decisions regarding the question(s) they are interested in answering, what variables they need to include, how "precise" their model needs to be, and how it corresponds to the elements of the situation.	Scientific models are abstracted schemas involving entities and relationships, meant to be useful across a range of particular circumstances. Procedures within the model space can be carried out to support inferences about the situation beyond what is immediately observable.	Like scientists, students of science should be familiar with the processes that lead to the development of scientific theories and the situated use of scientific models. Model elaboration, in which existing models are combined or extended to incorporate new data or to increase theoretical parsimony, is one such aspect of scientific inquiry: The user extends adapts, and connects models as prompted by the target situation.	Scientists reason through problems both as qualitative or physical relationships and as symbolic systems. This ability to articulate across multiple qualitative and/or quantitative representations or physical realities is crucial to students' development of scientific knowledge and ability.
Focal KSAs (Note: "ability" here means capability to reason as described in a given context with given models. No claim is made for "abilities" as decoupled from particular models.)	Ability to pose relevant questions about system to construct model Ability to relate elements of model to features of situation and vice versa Ability to describe (i.e., narrate) the situation through the entities and relationships of the model Ability to identify which aspects of the situation to address and which to omit. Decision-making regarding scope and grain-size of model, as appropriate to the intended use of the model.	Ability to reason through the concepts and relationships of a given model to make explanations, predictions and conjectures ▪ Qualitative reasoning through the model ▪ Quantitative reasoning through the symbolic representations associated with the model	Ability to identify links between similar models (that share objects, processes, or states) Ability to link models at different levels or focusing on different aspects of phenomena to create a larger, more encompassing model.	Ability to articulate meanings between qualitative and/or quantitative systems associated with scientific phenomenon. Ability to transform information between qualitative and/or quantitative systems associated with scientific phenomenon.
Characteristic features	Specific situation or data (either provided or previously generated by student), to be modeled, for some purpose. Correspondence must be established between elements of the model and elements of the situation.	Real-world situation and one or more models appropriate to the situation. Focus is on reasoning through the schema and relationships embedded in the model. Reasoning is as if the model is appropriate to the situation is the focus.	Real-world situation and one or more models appropriate to the situation, for which details of correspondence need to be fleshed out. Addresses correspondence between situation and models, and models with one another. Problem solution involves combining or making additions to existing models by, for example, embedding a model in a larger system, adding more parts to the model, or incorporating additional information about a real-world situation into the schema the model represents.	Multiple inter-related representation systems Task addresses relationship in expressions from one system to another

	Model Formation	**Model Use**	**Model Elaboration**	**Model Articulation**
Add'l KSAs	Familiarity with real-world situation Knowledge of model at issue Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) (e.g., STELLA, Marshall's arithmetic schema interface) Familiarity with required symbolic representations associated procedures (e.g., Marshall's schema forms, mathematical notation) Familiarity with task type (e.g., materials, protocols, expectations)	Familiarity with real-world situation Knowledge of model at issue Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) Familiarity with required symbolic representations associated procedures Familiarity with task type (e.g., materials, protocols, expectations)	Familiarity with real-world situation Knowledge of model at issue Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) Familiarity with required symbolic representations associated procedures Familiarity with task type (e.g., materials, protocols, expectations)	Knowledge of and ability to reason within qualitative and quantitative systems implied in the task. That is, this DP isolates the ability to move *between* systems, and therefore it presupposes students' ability to operate *within* the symbolic etc. systems involved. Knowledge of model at issue Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) Familiarity with required symbolic representations associated procedures Familiarity with task type (e.g., materials, protocols, expectations)

	Model Formation	Model Use	Model Elaboration	Model Articulation
Variable features	Is problem context familiar (i.e., degree of transfer required)? To what degree is the model prompted? Is model formation isolated, or in the context of a larger investigation? Complexity of problem situation; e.g., simplified situation vs. more factors and realism; simple mapping of word problem quantities, as in Marshall's schemas for arithmetic, vs. problem requiring large, hierarchical, and/or sequential modeling Complexity of the model; i.e., number of variables, complexity of variable relations, number of representations required, whether the model is runnable) Well-defined problem vs. ill-defined problem (or gradations thereof) Is extraneous information provided (makes tasks more difficult)? Kind of model needed for problem goal: simple & quick versus more exact and complex Role/depth of approximation required Degree of scaffolding provided Group or individual work?	Is problem context familiar (i.e., degree of transfer required)? Is model use isolated, or in the context of a larger investigation? Complexity of model Complexity of situation Complexity of reasoning required (e.g., number of variables in model, number of steps required) Model provided or generated by student? Data provided or generated by student? Degree of scaffolding provided (especially if model use involves strategies, alternate approaches, and multiple steps) Group or individual work?	Is problem context familiar (i.e., degree of transfer required)? Model given to student(s), vs. model to elaborate produced by student(s) themselves Complexity of elaboration required; e.g., minor modification of familiar model, vs. nesting of models, vs. elaboration to new previously unknown model. Is model use isolated, or in the context of a larger investigation? For example, must experimental work or supporting research be carried out in order to ground the elaboration? Single model to elaborate vs. establishing correspondence among models at different levels or with different foci? Degree of scaffolding provided (e.g., is need for elaboration prompted? Are hints or checklist provided to guide elaboration?) Group or individual work?	Articulation between semantic and symbolic systems, among different symbolic systems, or across multiple systems? Is problem context familiar (i.e., degree of transfer required)? Number of systems used (model, symbolic, physical, abstract) Complexity of systems Complexity of mappings (conditions, # issues to simultaneously consider) Prior exposure to representations and mapping conventions Is articulation the focus of a task, or is it part of a larger task? If part of a larger task, is the articulation problem presented to the student or is the need for, and carrying out of, articulation to be sought in the trace of a free-form solution trace? Degree of scaffolding provided (e.g., is need for elaboration prompted? Are hints or checklist provided to guide elaboration?) Group or individual work?

	Model Formation	Model Use	Model Elaboration	Model Articulation
Potential work products	Final version of model Explanation of model interpretation (as written or typed log or response, or audio transcript) Correspondence mapping between elements or relationships of model and real-world situation when there are gaps in one or the other to be filled Notes taken during model building process Trace of steps taken to build model See Scalise and Gifford (2006) for a taxonomy and illustrations of work product formats that are amenable to automated scoring.	Selection of hypotheses, predictions, retrodictions, explanations, and/or missing elements of real world situation Constructed hypotheses, predictions, retrodictions, explanations, and/or missing elements of real world situation, via ▪ Creation of one or more representational forms. ▪ Filling in given, possibly partially filled in, representational forms. Intermediate products developed in selection/construction of hypotheses, predictions, explanations, and/or missing elements Written/oral explanation of the hypotheses, predictions, explanations, and/or missing elements. Trace of actions taken in solution. Talk-aloud of solution. Critique of a given solution	Correspondence mapping between elements or relationships of model and real-world situation Correspondence mapping between elements or relationships of overlapping models. Final elaborated model (Physical, symbolic, verbal, etc., as appropriate) Trace of steps and provisional models Written/oral explanation of reasoning behind elaboration	Re-expression of information in one or more systems in terms of another system Cross-system problem solutions with mappings (e.g., force diagrams and equations). Can be prompted with "show your work." Verbal descriptions and explanations of meanings across representational systems Predictions for one system given information about an associated system Selection of system for scenario presented in terms of other systems

	Model Formation	**Model Use**	**Model Elaboration**	**Model Articulation**
Potential observations	Qualities of final model • Accuracy of determination of what variables are important to include in the model • Accuracy of representation of relations among variables • If runnable model, quality of model output • Appropriateness of degree of precision of model of phenomenon or system being modeled. Modeling process • Efficiency in terms of tools and representations • Quality (relevancy, accuracy) of questions asked about the system to inform the construction of a model. includes domain-specific heuristics and domain-specific explanatory schemas when these are targets of inference. • If talk-aloud, degree to which student talks about the meaning of the data • Quality of rationale student provides for steps in construction of model. Includes domain-specific heuristics and domain-specific explanatory schemas when these are targets of inference • Quality of rationale for what entities and relationships are expressed in the model, versus those which are omitted • Speed at which student forms model [indicates automaticity; Kalyuga, 2006]	Quality of students' explanations, predictions, or retrodictions as reasoned through the model ; e.g., correctness, appropriateness (i.e., quality of the product of model us). Qualities of solution procedure, such as appropriateness, efficiency, systematicity, quality of strategy, and effectiveness of procedures (i.e., qualities of the student's process). Quality of student's explanation of her own solution through a model (i.e., quality of the student's explanation of their process of model use, as distinct from the quality of the product of their reasoning).	Extent to which student catenates models appropriately across levels; that is, common entities and processes match up appropriately (e.g., individual-level and species-level models in transmission genetics) Quality of student explanation of modifications, in terms of features of data/purpose that require reasoning across levels/submodels Accuracy and completeness of mappings between a real-world situation and elaborated model.	Quality of operations applied *across* systems Extent to which student accurately maps one system into another, rather than back onto itself Accuracy of predictions in system y based on expressions in system x Accuracy and completeness of creation of system y based on expressions in system x Quality/appropriateness of description of meaning of information across systems Accuracy of selection of system (given example – i.e., instruction would have made the various systems explicit to students).
Selected References	diSessa (1993). Hunt & Minstrell (1994) Kintsch & Greeno (1985) Kintsch (1994) Kindfield (1999) Redish (204)	Stewart, J., & Hafner, R. (1994). Johnson-Laird (1983) Gentner (1983) Hestenes, Wells, & Swackhamer (1992)	Marshall (1993, 1995). Stewart & Hafner (1994). White & Frederiksen (1998).	Greeno (1989) diSessa (1983, 1993) Marshall (1993)

	Model Evaluation	Model Revision	Model-Based Inquiry
Summary	This design pattern supports developing tasks in which students evaluate the correspondence between a model and its real-world counterparts, with emphasis on anomalies and important features not accounted for in the model.	This design pattern supports developing tasks in which students revise a model in situations where a given model does not adequately fit the situation or is not sufficient to solve the problems at hand.	This design pattern supports developing tasks in which students work interactively between physical realities and models, using principles, knowledge and strategies that span all aspects and variations of model-based reasoning.
Rationale	It is essential to be able to determine the degree to which a model is in fact an appropriate method for reasoning about a physical situation. Examining evidence of the nature and quality of model misfit to data is important in this regard. Key to this endeavor is the degree to which predications and inferences from the model, given some of the data, are consistent with other parts of the data or further data that could be gathered.	Model-based reasoning concerns making inferences about real-world situations through the entities and structures of a model. When the model is not appropriate for the job at hand, either because it does not fit or it does not adequately capture the salient aspects of the situation, it is necessary to be able to revise the model.	Coordinating the aspects of model-based reasoning in inquiry requires not only being able to reason in each of the specific aspects, but to coordinate their use in iteratively building and testing models as inquiry proceeds. Metacognitive aspects of strategies for model-based reasoning are involved.
Focal KSAs (Note: "ability" here means capability to reason as described in a given context with given models. No claim is made for "abilities" as decoupled from particular models.)	In broad terms, ability to determine the appropriateness of a model for reasoning about a situation, for a given purpose. More specifically, ability to identify salient features of available data for comparison, and detect anomalies that available models cannot explain.	Ability, in a given situation, to modify a given model so that its features better match the features of that situation for the purpose at hand. More specifically: ▪ Recognizing the need to revise a provisional model. ▪ Modifying it appropriately and efficiently. ▪ Justifying the revisions in terms of the inadequacies of the provisional model.	Ability to carry out aspects of model-based reasoning when appropriate in an investigation, moving from one to another and using the results of each step to guide the next. Ability to monitor progress and results in inquiry cycle investigations (metacognition, self-regulation) Ability to take appropriate action with regard to model-based inferences in light of real-world feedback
Characteristic features	A model is proposed for a situation, and its suitability must be evaluated—is it satisfactory for the purpose? Where and how might it not appear to be adequate?	A situation to be modeled, a provisional model that in inadequate in some way, and the opportunity to revise the model in a way that improves the fit	Necessity to probe a problem by invoking or revising models to explain phenomena or make predictions. One or more inquiry cycles, and more than one aspect of inquiry, is required, so that managing interaction among them is required.

	Model Evaluation	Model Revision	Model-Based Inquiry
Add'l KSAs	Familiarity with real-world situation Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) Familiarity with required symbolic representations associated procedures (especially statistical methods) Familiarity with task type (e.g., materials, protocols, expectations) Familiarity with standards of quality & expectation in the field Ability to encode and represent evidence to be evaluated as an entity distinct from representations of the model	Ability to detect anomalies not explained by existing model (i.e., model evaluation) Familiarity with real-world situation Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) Familiarity with required symbolic representations associated procedures Familiarity with task type (e.g., materials, protocols, expectations) Ability to engage in model use	Familiarity with real-world situation Domain area knowledge (declarative, conceptual, and procedural) Familiarity with required modeling tool(s) Familiarity with required symbolic representations associated procedures Familiarity with task type (e.g., materials, protocols, expectations) Familiarity with standards of quality & expectation in the field Additional KSAs as might be required in particular aspects of model-based reasoning addressed in design patterns for those aspects
Variable features	Is the model-to-be-evaluated given, or was it developed by the student in the course of an investigation? Does the situation itself provide feedback about a model (e.g., as in interactive tasks such as troubleshooting) Is the model satisfactory or not satisfactory? If the model is not satisfactory, in what way(s) is this so? (E.g., lack of fit to observations, inappropriateness to project goal, wrong grainsize or aspects of phenomenon) Is problem context familiar (i.e., degree of transfer required)? To what degree is the model evaluation prompted? Complexity of problem situation Complexity of the model; i.e., number of variables, complexity of variable relations, number of representations required, whether the model is runnable) Is extraneous information present (makes tasks more difficult, because it evokes the need to evaluate whether certain aspects of the situation should not be modeled)? Group or individual work?	Is the model-to-be-revised given, or was it developed by the student in the course of an investigation? In what way is the model unsatisfactory: Lack of fit to observations, inappropriateness to project goal, wrong grainsize or aspects of phenomenon? Are the unsatisfactory aspects provided to the student, or to be discovered through model evaluation? Is model revision iterative, with feedback? To what degree is the model revision prompted? Is problem context familiar? Complexity of problem situation Complexity of the model; i.e., number of variables, complexity of variable relations, number of representations required, whether the model is runnable) Group or individual work?	Amount of scaffolding provided for working through inquiry cycles (from being walked through cycles, to support within cycles, to unsupported) Note: Scaffolding can be in the form of structured work products. Single cycle or multiple cycles anticipated? Problem given or self-determined in accordance with some criteria? Is problem context familiar? Time scale (e.g., relatively short snippets of fuller investigations as might appear in a large-scale test; short hands-on or simulation tasks, say 30-90 minutes; investigations that are carried out over days or weeks) Opportunities to engage in persuasion of peers in defense of their own solutions? Group or individual work? Variable features as might be required in particular aspects of model-based reasoning addressed in design patterns for those aspects

	Model Evaluation	**Model Revision**	**Model-Based Inquiry**
Potential work products	Verbal/written explanation of model-fitting actions [especially looking for prediction of new observations] Trace of actions Statements of hypotheses that motivate evaluation procedures Talk-aloud trace [which may exhibit evidence of model evaluation actions/reasoning] Representations and summaries of formal model-fitting tools such as statistical tests. Record of results of model-fit analyses on forms provided (note: this is a form of scaffolding) Explanation of results of model-fit analysis Record of hypotheses formulated and tested	Choice or production of revised model Explanation of reasoning for revised model Trace of models as constructed/revised (e.g., sequence of Genetics Construction Kit (GCK) models) Recordings or transcripts of what students said as they "thought aloud" while revising model Computer-kept records of inquiry steps in which model revision steps are embedded Notes written by students during model revision	Trace of models as constructed/revised during investigation Explanations of steps taken during an investigation, moving across inquiry steps. Audio-recordings or transcripts of what students said as they "thought aloud" while solving problems Computer-kept or notebook records of inquiry steps Filling out forms summarizing work by step or category. Note: Structured Work Products can be a form of scaffolding. Work products as might be required in particular aspects of model-based reasoning addressed in design patterns for those aspects Final and intermediate products of solution can hold indirect evidence about inquiry capabilities. Final products that illustrate appropriate models and conclusions suggest appropriate model formation and use at a minimum. Increasingly improved intermediate products suggest appropriate model evaluation and model revision or elaboration.
Potential observations	Comprehensiveness and appropriateness of methods of assessing model fit Systematicity of model evaluation procedures (e.g., are results of a test used to guide the choice of the next test?) Degree of integration of results from multiple tools/views for assessing model fits Quality of explanation of model fit Indication of which aspects of model do not fit, with respect to aspects of data and aspects of model Quality of determination of whether model misfit will degrade target inferences	Quality and appropriateness of model revisions in order to address inadequacies of provisional model. Degree of and appropriateness of general and/or domain-specific heuristics students use to revise their models. Quality of the basis on which students decide that a revised model is adequate Quality of explanation of the basis on which students decide that a revised model is adequate Efficiency of the process by which students evaluate existing models as deficient and revised models as adequate ,including use of optimal strategies, sequence, monitoring… This observable can be applied when model revision is part of a larger investigation. Extent to which students extract true results from a set of false models and recognize them as independent of the specific assumptions that vary across the (false) models.	Presence and quality of activity that reflects self-regulation, including ▪ Explanation of strategy ▪ Planning ▪ Reaction to feedback from other people or the situation itself Quality of student's explanation of her own solution through a model (i.e., quality of the student's explanation of their process of model use, as distinct from the quality of the product of their reasoning).
Selected References	Belsley, Kuh, & Welch (1980) Cartier (2000) Mosteller & Tukey (1977)	Mosteller & Tukey (1977) Rumelhart & Norman (1977) Stewart & Hafner (1991, 1994). Stewart, Hafner, Johnson, & Finkel (1992)	Mosteller & Tukey (1977) Stewart, Hafner, Johnson, & Finkel (1992). White & Frederiksen (1998)

!

Sponsor
The National Science Foundation, Grant DRL - 0733172

Prime Grantee
SRI International. *Center for Technology in Learning*

Subgrantees
University of Maryland
Pearson