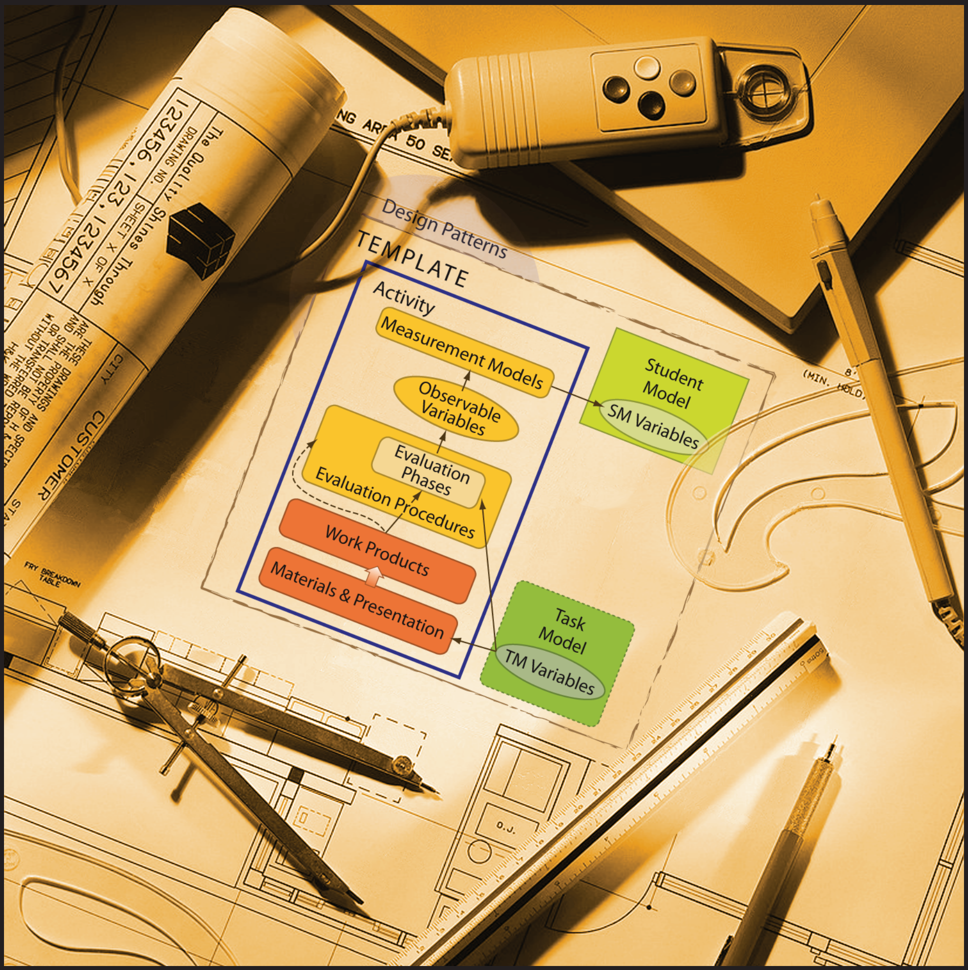


Large-Scale Assessment Technical Report 14|
August 2013

A "Wizard" for Authoring Scenario-Based Tasks, Using Evidence-Centered Design Principles and Structures

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

Robert Mislevy, Educational Testing Service
Patricia Schank, Mingyu Feng, Ron Fried, Britte Chang, Eric Snow, and Geneva Haertel, SRI International
Alexis Mitman Colker, SRI International Consultant

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*

DESIGN TO STATE LARGE-SCALE SCIENCE ASSESSMENT

TECHNICAL REPORT 14

A "Wizard" for Authoring Scenario-Based Tasks, Using Evidence-Centered Design Principles and Structures

August 2013

Prepared by:

Robert Mislevy – Educational Testing Service
Patricia Schank, Mingyu Feng, Ron Fried, Britte Cheng, Eric Snow, Geneva Haertel – SRI International
Alexis Mitman Colker – SRI International Consultant

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).

Table of Contents

ABSTRACT

This report describes how the PADI assessment design framework was used to build a "wizard" to author a family of scenario-based assessment tasks. . The tasks address standards in model-based reasoning and ecology, in the context of invasive species. . The wizard was structured around the underlying principles and PADI representations for evidence-centered design to help test developers write scenario tasks that were at once faithful to the standards and embodied coherent assessment arguments, while drawing upon their knowledge of the domain. .

1.0 Introduction and Overview

A software "wizard" is a tool that leads the user through a structured interview that helps a user achieve a goal. Familiar examples are TurboTax's wizard for completing income tax forms and Dell's wizard for choosing and customizing a personal computer. The choices and the user-supplied input are tuned to the user's knowledge and way of thinking. The structure of the interview and the processes behind the scenes can exploit sophisticated knowledge to carry out the necessary processes, hiding much of the complexity from the user.

This report describes the PADI Storyboard Wizard, a wizard that helps task developers create instances of scenario-based assessments that address standards concerning model-based reasoning and ecology, in the context of invasive species. The wizard was developed in the project "Application of Evidence-Centered Design to State Large-Scale Science Assessment," supported by the National Science Foundation (NSF) to apply the principles of evidence centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2003; Mislevy & Haertel, 2006) in the context of a state-level large-scale accountability assessment in science. It builds on the PADI assessment design system, developed in a previous NSF project called Principled Assessment Design for Inquiry.

Both PADI and the current project develop support structures for assessment design, with special attention on science assessments and particularly ones that are more complex in the sense of being scenario-based or interactive. The current project additionally focuses on assessments that are feasible for large-scale assessments such as state accountability tests. The wizard described here, for example, is grounded in the science standards of the state of Minnesota, and produces tasks patterned after the computer-based scenario tasks of the science tests of the Minnesota Comprehensive Assessment (specifically, MCA II). Each task generated by working through the wizard is a scenario-based task requires examining the changes in an ecosystem caused by the introduction of an invasive species and revising a model of the food web for the ecosystem accordingly. There are hundreds of invasive species in North America, which may not be so good for the ecology but it means the wizard can help users create hundreds of unique scenario-based tasks.

Here are the key ideas behind the wizard: Designing scenario-based, multiple-scene, multiple-item tasks that address an assigned set of benchmarks derived from standards is a significant challenge, even to experienced test developers. There are several design constraints that must be met simultaneously, including items addressing each of the targeted benchmarks, sound science content, valid assessment argument structures for each item, and a coherent narrative structure connecting the scenes. To create a wizard of the kind described here, a team of expert developers first creates an essential structure for a family of task variants to be fleshed out with particular contents and contexts. They draw upon substantial design experience, content knowledge, and ECD framework. This structure addresses difficult aspects of the joint constraint satisfaction problem. A wizard is then built around this structure. The wizard walks a local author (who is sufficiently familiar with science content) through a process that creates a specific instance of the task family that suits the local needs of the assessment and the students of interest.

Sections 2 and 3 provide background for the wizard development. Section 2 briefly reviews ECD and PADI. Section 3 discusses the foundational material the wizard is based on: the structure of

MCA scenario-based tasks, the standards being addressed, a model-revision design pattern, and the more specific science content for an invasive species family of tasks.

Section 4 looks more closely at the wizard itself. It contains an overview description, a walk through its screens, and an example of one form of the output.

Section 5 discusses the role of wizards more broadly, emphasizing the value of an overarching assessment design-and-delivery system such as PADI.

The purpose of developing this wizard is not to provide a specific tool to create a family of tasks. It is rather to demonstrate how a principled framework for assessment design (e.g., ECD) and supporting tools and representations (e.g., the PADI system) make efforts like this more efficient, the outcomes more valid, and the processes for doing so scalable. The expert-level thinking that goes into designing complex tasks for hard-to-assess proficiencies need not be the isolated in the heads of experts alone, but can be made more public through the representations and more accessible to a wider range of users when built into tools such as the wizard.

2.0 ECD and PADI

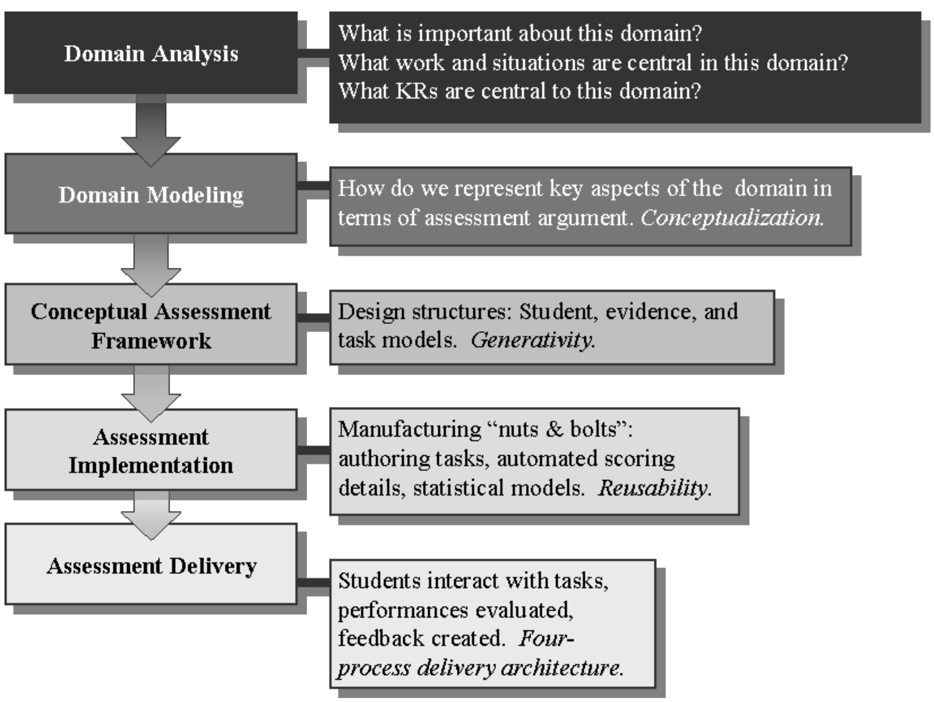
Evidence-centered assessment design (ECD) provides principles, patterns, and examples to guide the task designers through articulating the theoretical foundation to the operational work of assessment development (e.g., item writing, directions, test administration, scoring procedures) (Mislevy, Almond, & Lukas, 2004). This structured framework enables designers to more explicitly control the elements and underlying processes of assessment design, and contributes to both validity arguments and operational efficiencies. This section provides a brief overview of the key ideas of ECD, noting their roles in developing the wizard. The construction of the wizard used tools and representations from the PADI project (Baxter & Mislevy, 2005; Mislevy & Riconscente, 2006).

Two complementary ideas underlie ECD. The first is an overarching conception of an assessment as an argument from imperfect evidence. Messick (1994) lays out the basic narrative, saying that

> [We] would 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)

The second idea is distinguishing the five layers shown in Figure 1, which conceptualize different kinds of work carried out in the design and implementation of an assessment.

Figure 1: ECD layers



2.1 Domain Analysis

Domain analysis concerns marshaling substantive information about the domain. It leads us to understand the knowledge, skills, and abilities people use in a domain of interest, the representational forms they use, characteristics of good work, and features of situations where this knowledge is used. Practical experience, standards documents, and research studies are examples of sources of such information. It is all relevant to assessment design, but not necessarily organized in terms of assessment arguments or assessment practices.

The wizard at issue was developed in the context of the state of Minnesota's science content standards for middle school, so important sources of information for developing the wizard included the *Minnesota Academic Standards Science K-12*[1] and the *Minnesota Comprehensive Assessments-Series III (MCA-III) Test Specifications for Science*[2]. The wizard also draws on research in model-based reasoning, using the design pattern for model revision discussed below, and on substantive research about invasive species, food webs, and adaptations of organisms to ecological systems. Section 3 will provide additional details about the standards and substantive grounding for the wizard.

2.2 Domain Modeling

In Domain Modeling, information identified in Domain Analysis is organized along the lines of assessment arguments. Supporting tools such as Toulmin diagrams (Mislevy 2003, 2006) and design patterns (Liu & Haertel, 2011; Mislevy et al., 2003) help developers think through the assessment argument without getting tangled up in the details of implementation.

Design patterns in particular lay out a design space to help task developers create tasks that embody a coherent assessment argument. Table 1 presents the main attributes of a design pattern, and relates them to the elements of an assessment argument. (KSA stands for "knowledge, skill, and/or ability" – whatever the capability of interest for assessing is. KSA is a broad term, meant to encompass proficiencies that could be conceived under any psychological perspective—behavioral, trait, information-processing, situative—and could be long-standing or susceptible to change over long or short intervals. The structure of a design pattern is agnostic as to grain-size and psychological perspective.)

Generative schemas for families of tasks are especially helpful for assessments that need to generate multiple forms, address hard-to-measure skills, or use unfamiliar forms of assessment. The PADI project focused on important but hard-to-assess aspects of inquiry in science, such as experimental and observational studies, model-based reasoning, and systems thinking.[3]

The wizard addresses both a hard-to-assess skill (model revision) and uses an unfamiliar task type. It based on the multiple scene, technology-enhanced, scenario tasks in the Minnesota Comprehensive Assessment. Section 3 will look more closely at the particular design pattern the

---

[1] Accessed August 8, 2013 from http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=005263&RevisionSelectionMethod=latestReleased&Rendition=primary

[2] Accessed August 8, 2013 from http://education.state.mn.us/mdeprod/idcplg?IdcService=GET_FILE&dDocName=003177&RevisionSelectionMethod=latestReleased&Rendition=primary

[3] PADI research reports presenting design patterns appear at http://ecd.sri.com/publications.html

wizard draws upon. The key idea, however, is this: The expert team that developed the wizard relied on the information in the design pattern to create a task structure that reflected a coherent, valid assessment argument; they used their substantive knowledge to indicate content and relationships in ecology and science reasoning to lay out schemas for tasks that would address the targeted standards; and they used the forms of the MCA to frame the contents and flow of the tasks that would result.

Thus, a task developer using the wizard to develop a task for revising a food-web model would have used their familiarity with their students and the particular content they chose to build a sophisticated scenario-based task that incorporated the expert team's knowledge of the domain, the standards, and assessment design theory. The wizard blends expert knowledge and local knowledge to produce tasks with high quality that are tuned to users' needs.

Table 1: Key Attributes of a Design Pattern

Attribute	Definition	Assessment Argument Component
Rationale	Nature of the KSA of interest and how it is manifest	Warrant
Focal KSA	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 construct relevant; Alternative Explanation, if construct 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 KSA	Data concerning students' actions
Characteristic Features	Aspects of assessment situations 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 KSA.	Data concerning situation

2.3 Conceptual Assessment Framework (CAF)

The Conceptual Assessment Framework (CAF) concerns technical specifications for operational elements. An assessment argument is now expressed in terms of coordinated pieces of machinery such as measurement models, scoring methods, and delivery requirements.

The wizard walks a test developer through a series of steps to gather information and materials to author a food-web revision task, focusing on the substance of the problem as it is seen from the developer's perspective. The developer is not necessarily familiar with the more technical data structures that may be needed to structure this information for use in a given assessment program.

A wizard not only structures the developer's assembling of local knowledge, but also combines it with the knowledge built in to the wizard to produce output that formats the results in a way they are most useful. PADI design system templates (Riconscente, Mislevy, Hamel, & PADI Research Group, 2005). are used to structure output in the wizard.

2.4 Assessment Implementation

The Assessment Implementation layer encompasses (possibly ongoing) activities that prepare for operational administration, such as authoring tasks, calibrating psychometric models, piloting and finalizing evaluation procedures, and producing assessment materials and presentation environments.

The wizard is quite specifically a tool to facilitate a key facet of assessment implementation, namely task authoring. Note that while it resides in the assessment implementation layer, it draws on insights from the previous layers and looks ahead to use in the following layer. The deep interconnections among thinking in the distinct layers is a hallmark of evidence-centered design, and a wizard uses the deep relationships in ECD in a way that developers can benefit from them without having to become experts in them themselves.

2.5 Assessment Delivery

Assessment Delivery addresses the processes of presenting tasks to examinees, evaluating performances to assign scores, and reporting the results to provide feedback or support decision making (Almond, Steinberg, & Mislevy, 2002). The four-process architecture is viewed in terms of processes and messages whose meaning is grounded in Domain Modeling, whose structure is laid out in the CAF, and whose pieces are built in Assessment Implementation.

Different forms of assessment objects are required in different assessment settings. For example, the assessment objects needed for a paper-and-pencil form differ notably from a computer-administered interactive task, and from a set of notes to guide a teacher through an in-class collaborative task – even if the same information is required in all cases. The output of a wizard needs to be constructed in knowledge of the delivery system that its product is to be used for. The wizard described in this report puts out a summary form that would be provided to a test implementer in the MCA system.

Alternative formatting of output could be provided, without requiring extra work on the part of the wizard user. For example, output could include a rendering for high-quality printed form of the task and data files structured as needed for presentation and scoring in a computer-based delivery system. Knowing the data structures of the files needed for presentation, user interaction, and scoring that are required, a wizard's programmers can assure that the output of the user-friendly interactions will produce the technical elements needed in even quite complicated interactive, computer-based tasks. Wizards of this type using ECD include the PADI Mystery Powders wizard (Hamel, Mislevy,& Winters, 2008; Seibert et al., 2006) and the assessment authoring interface of the Cisco Networking Academy's Packet Tracer tool for interactive simulation-based tasks of computer network troubleshooting tasks (Frezzo, Behrens, & Mislevy, 2009).

3.0 Foundations of the Wizard

Creating a complex assessment task is an exercise in design under constraints. A test developer must work to achieve assessment goals under a wide variety of constraints—logistical, psychometric, and substantive. This is a difficult challenge, even for experienced test developers, when the form, substance, or goals of the intended task are not familiar (Fulkerson, Nichols, & Mittelholtz, 2010; Nichols & Fulkerson, 2010). A successful wizard is not simply about obtaining locally pertinent information from the user; it is about having solved some of the difficult problems with conflicting and hard-to-meet constraints in hard problems. This section looks more closely at the knowledge about the structures, standards, science substance, and assessment-design theory that are built into the current wizard.

3.1 MCA Scenario-Based Tasks

The structure of the tasks that the wizard supports is based on the scenario tasks of the Minnesota Comprehensive Assessment in science. The MCA Science test consists of computer-delivered scenario-based assessment tasks. Development of these tasks begins with the writing of storyboards, which serve as contexts for standards-aligned items. There are four to six scenes in a scenario. Each can present information to a student, in the form of text, graphics, audio or video clips, a simulation space to act in, or some combination of these. Most also present an "item" based on information from the current and previous screens. The items range from familiar multiple-choice items to figural response items, graphical 'hot spot' items, drag-and-drop, filling in knowledge representations such as graphs, and typed-in open-ended responses. As this is written, examples of scenario-based tasks are available for inspection at http://www.mnstateassessments.org/item-samplers/#onlineItem.

Each MCA scenario-based science task is designed to address a specified set of Minnesota standards (*Minnesota Academic Standards Science K-12*). The sets are determined by which standards are to be addressed in a given assessment year, and reflect both the importance of the standards and which standards have been addressed in previous years (*Minnesota Comprehensive Assessments-Series III (MCA-III) Test Specifications for Science*). Test developers thus receive a writing assignment that consists of a set of standards to be addressed in a scenario-based task, usually each by an item on one of the scenario scenes. Additional standards may be addressed if the developers find they fit naturally into the scenario, given that the assigned standards have all been addressed.

In practice, the scenario-based tasks are created in two stages. The first is writing a storyboard, which contains drafts of the scenes and stimulus materials. The storyboard is intended to be a satisfactory base for writing items that address the assigned standards. Storyboards undergo several reviews before moving to the second stage, writing the individual items for the scenes. The storyboard reviews examine issues such as appropriateness, science content accuracy, feasibility to implement, and sensitivity. When a storyboard has passed reviews, perhaps with modifications, test developers who may be the same or different as the storyboard writers then write the individual items so as to address the assigned standards.

The wizard addresses the standards shown in Table 2, as amplified by the accompanying benchmarks. Simply addressing the standards is a long way from knowing how to create a storyboard and tasks, though. We turn next to a design pattern that supports understanding of how to create a storyboard that presents a coherent thematic set of scenes, which in turn supports creation of items to assess a key but hard-to-assess aspect of science inquiry, namely model revision.

Table 2: Standards and Benchmarks Addressed in Wizard

Interdependence among Living Systems

Standard: The interrelationship and interdependence of organisms generate dynamic biological communities in ecosystems.

Benchmark 9.4.2.1.1: Describe factors that affect the carrying capacity of an ecosystem and relate these to population growth.

Item Specifications

- Examples of factors include food or nutrient availability, predation, competition, population density, disease and waste removal
- Contexts will use examples of Minnesota ecosystems when appropriate

Benchmark 9.4.2.1.2:.Explain how ecosystems can change as a result of the introduction of one or more new species. For example: The effect of migration, localized evolution or disease organisms.

Item Specifications

- Contexts for items will use examples of Minnesota ecosystems when appropriate
- Items may require students to predict, analyze and reflect on global issues
- Items may include invasive species

Standard: The interrelationship and interdependence of organisms generate dynamic biological communities in ecosystems.

Benchmark 9.4.3.3.5: Explain how competition for finite resources and the changing environment promotes natural selection on offspring survival, depending on whether the offspring have characteristics that are advantageous or disadvantageous in the new environment.

Item Specifications

- Contexts for items will use examples of Minnesota ecosystems when appropriate

Human Interactions with Living Systems

Standard: Human activity has consequences on living organisms and ecosystems.

Benchmark 9.4.4.1.2: Describe the social, economic and ecological risks and benefits of changing a natural ecosystem as a result of human activity.

Item Specifications

- Contexts for items will use examples of Minnesota ecosystems when appropriate

Nature of Science and Engineering

Standard: Science is a way of knowing about the natural world and is characterized by empirical criteria, logical argument and skeptical review.

Benchmark 9.1.1.1.7: Explain how scientific and technological innovations—as well as new evidence—can challenge portions of, or entire accepted theories and models including, but not limited to: cell theory, atomic theory, theory of evolution, plate tectonic theory, germ theory of disease, and the big bang theory.

Item Specifications

- Items will address theories, models and the validity of scientific knowledge in the context of life science

Standard: Scientific inquiry uses multiple interrelated processes to investigate and explain the natural world.

Benchmark 9.1.1.2.2: Evaluate the explanations proposed by others by examining and comparing evidence, identifying faulty reasoning, pointing out statements that go beyond the scientifically acceptable evidence, and suggesting alternative scientific explanations.

Item Specifications

- Items may require students to evaluate a set of data to formulate possible conclusions

3.2 *The Design Pattern for Model Revision*

The PADI project and the present follow-on project developed a large number of design patterns to support task design for aspects of science inquiry. Among them is a suite of design patterns for model-based reasoning, including Model Formation, Model Use, Model Elaboration, Model Articulation, Model Evaluation, Model Revision, and Model-Based Inquiry (Mislevy, Riconscente, & Rutstein, 2009). The wizard uses the design pattern for model revision, summarized in Table 3.[4] This section looks more closely at this design pattern, as it helps bring out design decisions that are built into the wizard.

Rationale. Model-based reasoning is a complex skill that is carried out in some context, with some model(s). There are a number of distinguishable (though deeply interrelated) aspects of model-based reasoning, which can be assessed individually or in concert. Model revision is one of them. As seen in the following discussion, there are many ways and many kinds of tasks in which assessing students' capabilities with model revision can take place. The design pattern is meant to provide support for test developers thinking about how to assess this inquiry skill, giving them advice about issues they ought to consider and options for building tasks.

In tasks produced though the wizard, the model at issue is a food web for an ecosystem that has recently changed because an invasive species has been introduced.

Focal KSA. The capability at issue in model revision is modifying a given model so that its features better match the features of a situation for the purpose at hand. More specific aspects that can be identified and addressed in assessment are 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.

While other standards in addition to inquiry are addressed in the tasks produced by the wizard, one or more of the items will tap whether the student recognizes the inadequacies of the original food web and modifies it in light of what is known about the invasive species.

Additional KSAs. Additional KSAs are knowledge and skills that can be required in a task in addition to the Focal KSAs. Depending on the contextualization of the task in use, the intended students to assess, and the intent of measurement, Additional KSAs can be either appropriate to include or not. A design pattern does not know this; these are design decisions the test developer must make in light of her local knowledge. For example, is it desired to assess model revision in the context of a familiar context with familiar content so that only the model revision issues are at issue? Or is it desired to jointly assess whether a student can carry out model revision with a model that is also at issue and the student may or may not understand? Is a provisional model to be presented, or should it arise from the students' own investigations?

In the wizard, it was decided that students would be presented with the initial model (the food web before the invasive species arrived) and shown data and provided information about the invasive species. This decision forgoes evidence about the students' investigative capabilities but increases the chances of obtaining good evidence about model revision. The content concerning ecology, including the meaning of the food web, is required to do well in the task; it is a potential target of inference as well as model revision per se. A test developer using the wizard can choose

[4] See Mislevy, Riconscente, & Rutstein, 2009, for fuller discussion, and for interactive on-line versions of all the model-based reasoning design patterns, see http://design-drk.padi.sri.com/padi/do/NodeAction?state=listNodes&NODE_TYPE=PARADIGM_TYPE

whether to use a familiar invasive species or an unfamiliar one, so as to stress to lesser or greater degrees reasoning with new biology content.

Potential Work Products. A design pattern offers a variety of ways one might get evidence about the KSAs, in order to help the task developer come up with good ways to obtain evidence under whatever constraints the assessment must operate. Constraints are tightest in traditional paper and pencil (P&P) testing which must be automatically scored, for example; looser if open-ended responses can be scored by hand, different yet again if the tasks are computer delivered and can capture drag-and-drop or figural responses, and very much open-ended for tasks that are used in informal classroom work.

The wizard directly supports the construction of P&P forms that are appropriate for classroom use and could optionally be modified to produce the same output in data structures tuned to presentation in a computer-based delivery system, such as was actually used in the MCA. The wizard allows the user to create both constrained response and open-ended tasks. It provides "item ideas" they can adapt directly to the invasive species content they have chosen to use, but allows for new items to be entered.

Potential Observations. Whereas work products address the form of student responses, potential observations is about what, conceptually, are the nuggets of evidence we want to look for in responses. For example, one of the potential observations is explanation of reasoning for the revised model. In an open-ended explanation, the developer must provide a rubric for a human or automated scoring routine to evaluate whatever form of work product has been specified to be scored in such a way as to determine the quality of explanation. In a multiple-choice test, alternative explanations could be provided, and the student would indicate the best one. In a semi-constrained response format, the student might construct an explanation from a series of drop-down menus. The work products are different, but in all cases the conceptual nature of the observation—the evidence about the student's thinking—addresses the same facet of knowledge. Different work products have different evidentiary properties, but the potential observations category of a design pattern helps ensure that whatever forms of task are used, to the best that can be done the evidence being provided is about the knowledge and skill that are the target of assessment.

Both the scene skeletons and the item ideas in the wizard have been structured so as to elicit good evidence about the standards being addressed in the task family. It should be clear that choosing task features, work product formats, and response modes that capture the right evidence is not always straightforward. The wizard builds in a great deal of expert-level design thinking for doing so, for the family of tasks it supports.

Characteristic Features. At the heart of assessing model revision is having the student(s) work in a situation where a provisional model is seen to be inadequate in some respect, and must be revised to account for anomalous data or observations.

The wizard scenes quite specifically address this characteristic of tasks designed through it. This is probably the most challenging aspect of model-revision tasks to get right, since it requires thinking across multiple time points in a process as opposed to simple encapsulated situations like multiple choice tasks. As such, this aspect of task construction is most challenging for users who are not familiar with inquiry assessment; and as such, the wizard provides the strongest support. The scenes are specifically designed to present the original model and bring out, as the student works through them, ways it is inadequate and ways it might need to be revised.

Variable Features. A task can have the characteristic features needed to elicit evidence about the Focal KSAs, yet vary in many ways that affect its difficulty, the aspects of knowledge and skill it emphasizes, and the logistical and operational constraints it must meet. "Variable features of tasks" is the attribute where a design pattern makes the test developer aware of some of the choices that are available.

The wizard builds in some choices and leaves others to the user:

- *Is the model-to-be-revised given, or was it developed by the student in the course of an investigation?* Given.
- *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?* The original food web is unsatisfactory because it doesn't account for the invasive species. Screens develop this information, and suggest the ways that the invasive species will be related to the original creatures in the food web.
- *Is model revision iterative, with feedback?* Not iterative.
- *To what degree is the model revision prompted?* Strongly prompted.
- *Is problem context familiar?* Up to the user, as they choose the invasive species and ecosystem the task will address. Could be very familiar, such as a local one they have been discussing in class, or completely unfamiliar (even a fictitious example on an alien planet) which would put more emphasis on reasoning through the ecological model.
- *Complexity of problem situation.* The general complexity level of the scenario is determined through the kinds of information suggested in each scene and the kinds of items each will support. However, with the option to specify stimulus materials and create items, the user can adjust the difficulty of the task overall and in each scene.
- *Complexity of the model (i.e., number of variables, complexity of variable relations, number of representations required, whether the model is runnable).* The food web is the essential model, so the user's choice of the invasive species and associated food web is key: The more complex the web, with more interactions and species, the more difficult the task will be within the bounds determined by the general problem situation. The model is not runnable in these tasks. (A more ambitious task family could be envisioned where the invasive species situation played out interactively over time, and more aspects of inquiry such as generating and testing hypotheses, and feedback effects related to systems thinking, could be incorporated.)
- *Group or individual work?* As configured, the wizard provides a framework for working through the scenes and addressing questions, in a P&P format. It could be used for either individual or group work in the classroom, at the discretion of a teacher, for example. If used for group work, it would be possible using the items and evaluation schemes to score the work of a group of students. No support is provided, however, for assessing individuals' interactions with group members.

Also left to the user are the specifics of decisions about the particular content of the problem: What invasive species will be addressed? Where is it in the food web (predator, prey, producer)? What is the ecosystem at issue? (Minnesota's Test Specifications emphasize local context for science tasks, so zebra mussels and wild parsnips would be better choices for the MCA than Burmese pythons and kudzu.)

Narrative Structures. Story arcs based on investigation, change over time, and cause and effect are well suited to eliciting evidence about model revision. The wizard uses change over time as its narrative structure, so that the initial model is inadequate because it doesn't account for the new invasive species in the ecosystem.

Table 3: Summary Form of a Design Pattern for Model Revision

Attribute	Attribute Values
Rationale	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.
Focal KSA	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.
Additional KSAs	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
Potential Work Products	Choice or production of revised model Explanation of reasoning for revised model Trace of models as constructed/revised. 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.
Potential Observations	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.
Characteristic Features	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.
Variable Features	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?
Narrative Structures*	Investigation; Change over time; Cause and effect

* Attribute added for tasks that require a story arc.

3.3 The Content Being Addressed

The expert team that created the wizard structure knew they needed to create a shell for model-revision tasks and it needed to provide opportunities for assessing specific content along the way (Section 3.2). They needed to come up with a scientifically sound schema for addressing these aspects of science proficiency that was also interesting to middle school science students, fit the MCA scenario-based task structure, and could support multiple unique tasks in a family. The schema they developed was based on revising the food web model for an ecosystem after an invasive species has been introduced.

The scenario developed by the expert team is based on a growing real world concern related to invasive species. Invasive species are flora or fauna that are non-native to an ecosystem and adversely affect the habits and bioregions they invade. The specific example used was of the Burmese Python, a snake native to Southeast Asia and one of the largest snake species on earth. This python is now known to be breeding in the Florida Everglades National Park and spreading throughout south Florida. Over 1,800 pythons have been removed from the park and surrounding areas since 2002—and that likely represents only a fraction of the total python population in the park.

The introduction of the Burmese Python was believed to be the result of both accidental and intentional release by pet owners. Their introduction to the Everglades National Park's ecosystem has had devastating consequences. Burmese pythons are top predators and have been found to feed on a wide variety of mammals and birds that are in the Everglades existing food web—even the occasional alligator! Since they prey on native wildlife, compete with other native predators, and are very prolific, the Burmese Pythons are seriously disrupting the park's (and even south Florida's) ecological environment, and further threatening many endangered native species.

The scenario, illustrated below, asks students to review and interpret data in order to describe how the introduction of the Burmese Python to Florida Everglades National Park's ecosystem has changed the existing food web and habitat. Because this scenario addresses both content knowledge (in ecology) and process skills (namely, model creation and revision), it provides a good framework for assessing science knowledge and proficiency.

4.0 A Closer Look at the Wizard

The PADI Storyboard Wizard scaffolds assessment developers in the creation of storyboards and ideas for assessment items that involve food webs and invasive species. The Storyboard Wizard was designed in collaboration with science subject matter experts at Pearson and SRI and designers at SRI and Codeguild, with input from assessment experts at Educational Testing Service (ETS) and SRI. The wizard was implemented in Wordpress, an open-source content management system.

The home page of the PADI Wizard is shown in Figure 2. The wizard walks an assessment developer through five scenes about the effects of some invasive species in an ecosystem (see Figure 3Figure 8). First, the assessment developer specifies the invasive species and ecosystem (Figure 3). The wizard then prompts the assessment developer to enter text and upload images to tailor the scenes to their invasive species, drawing on their knowledge of the ecosystem (Figure 4Figure 8). When the assessment developer has entered information and images for all five scenes, the system creates and presents their storyboard on one (long) page (Figure 9).

For each scene in the completed storyboard, the assessment developer can then add items that measure different aspects of science knowledge and skills (Figure 10Figure 13). The system offers item ideas that the developer can tailor to their invasive species to target the skills they are interested in assessing (Figure 13).

Figure 14 shows a complete version of a storyboard after working through the wizard and adding some items. It shows the information and representations the user provided in the process. Figure 15 shows a shorter, outline view of the completed scenes and KSAs addressed by each item, reached by clicking “Show Outline”. Alternative output formats can be student-ready paper and pencil sheets for working through the problem, or files in the formats that are required for computer delivery in an online assessment presentation system (e.g., by clicking “Export to XML”).

Figure 2: PADI Storyboard Wizard: Home page.

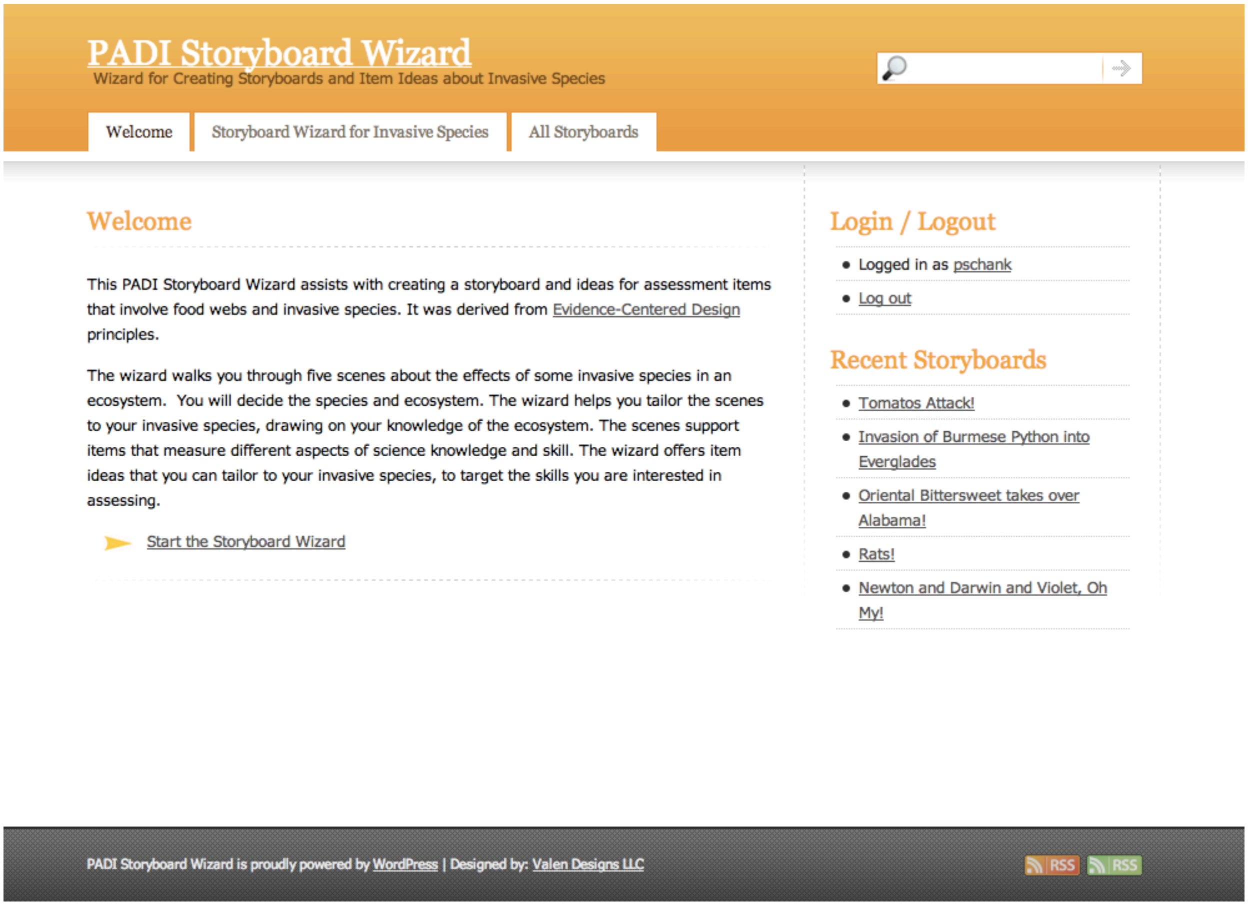


Figure 3: Storyboard Wizard Step 1: Describe the invasive species.

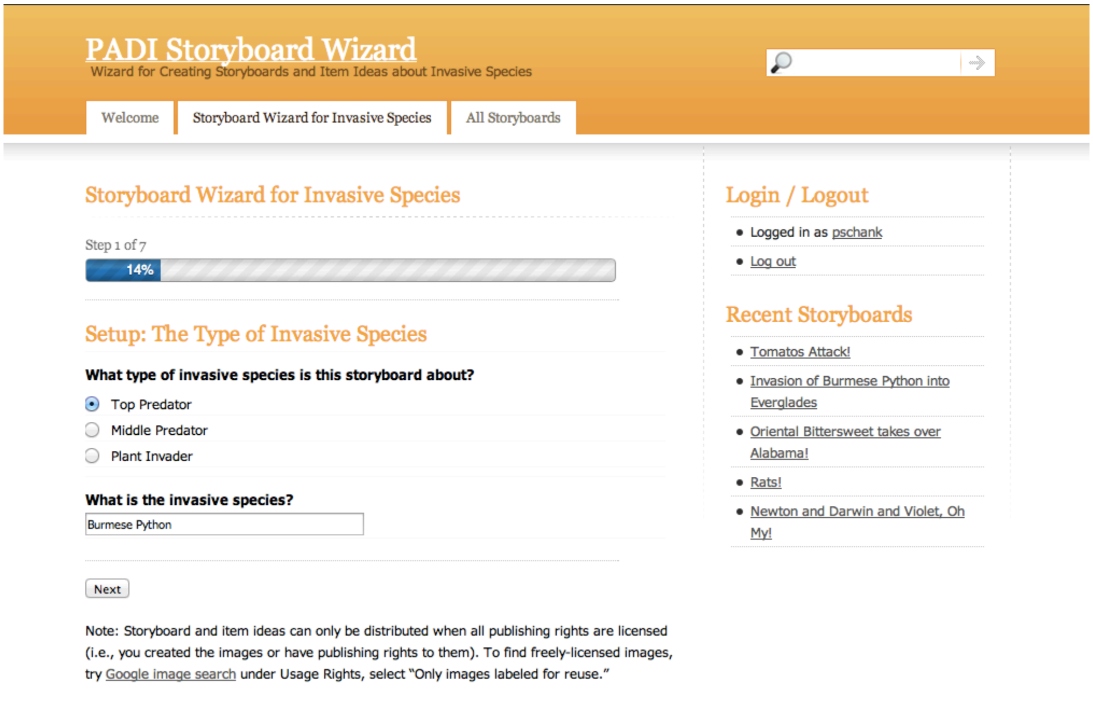


Figure 4: Storyboard Wizard Step 2: Describe the original food web, before the invasive species is introduced.

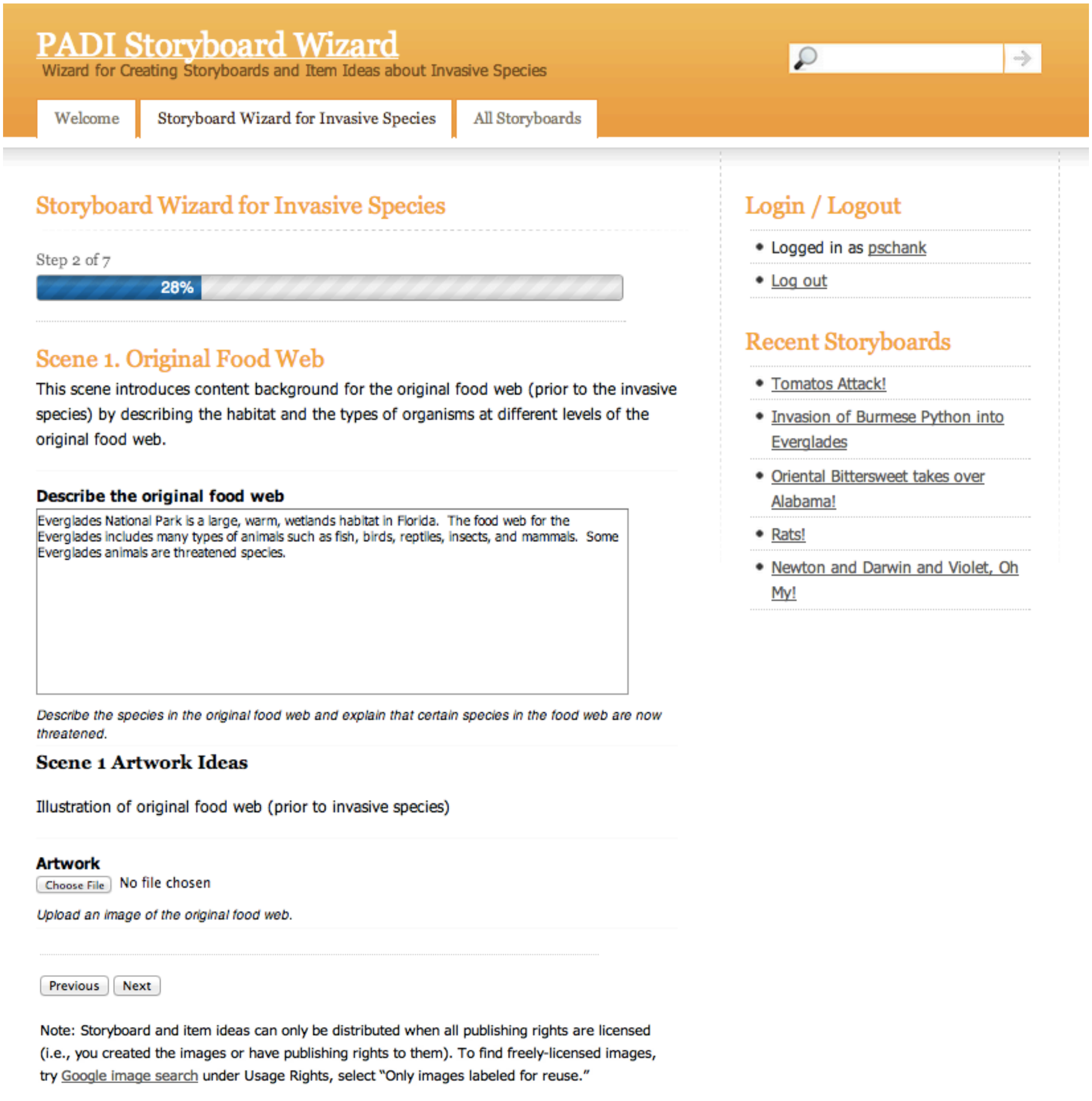


Figure 5: Storyboard Wizard Step 3: Describe how the invasive species entered into the food web.

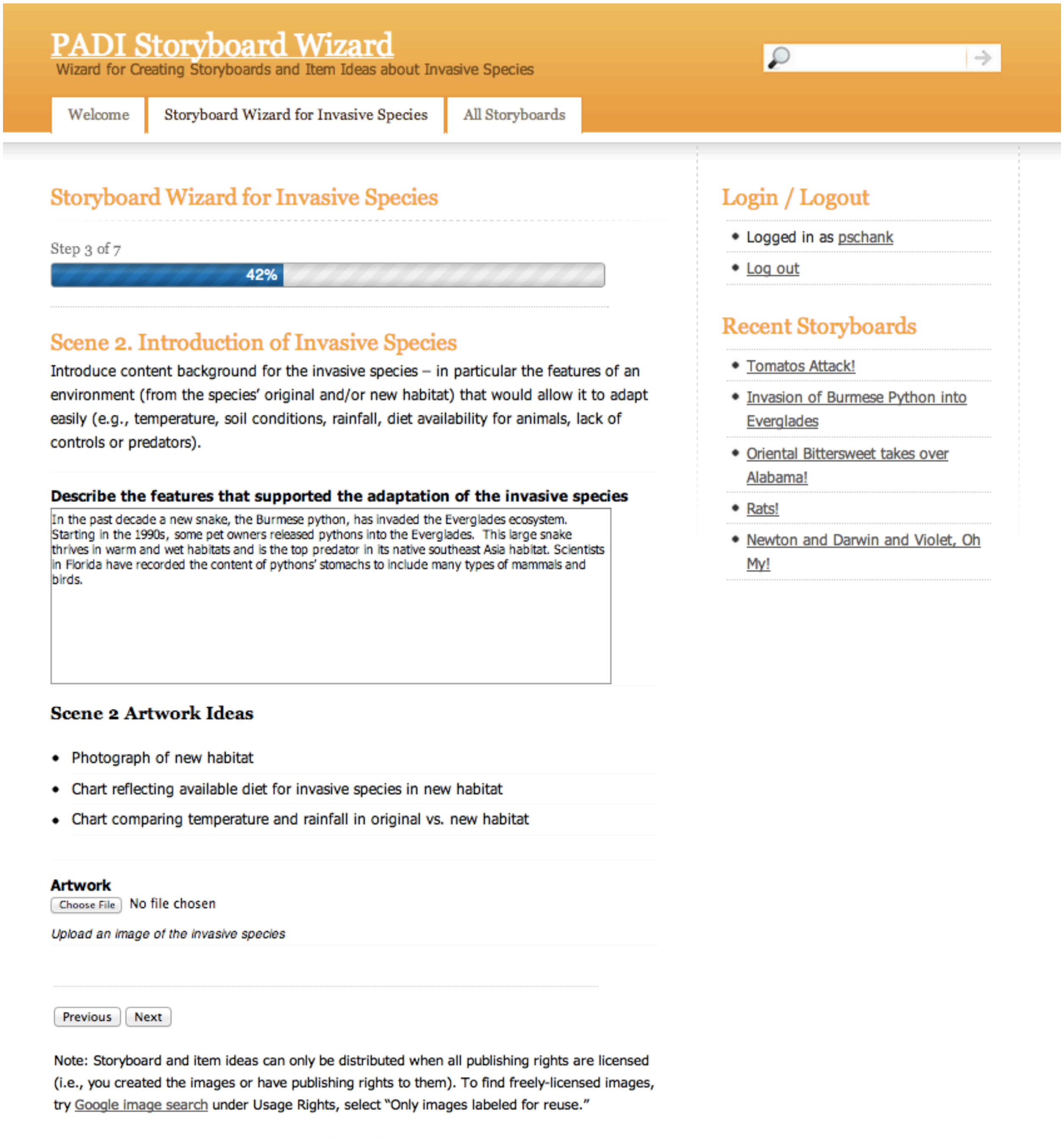


Figure 6: Storyboard Wizard Step 4: Describe the growth of the population of invasive species.

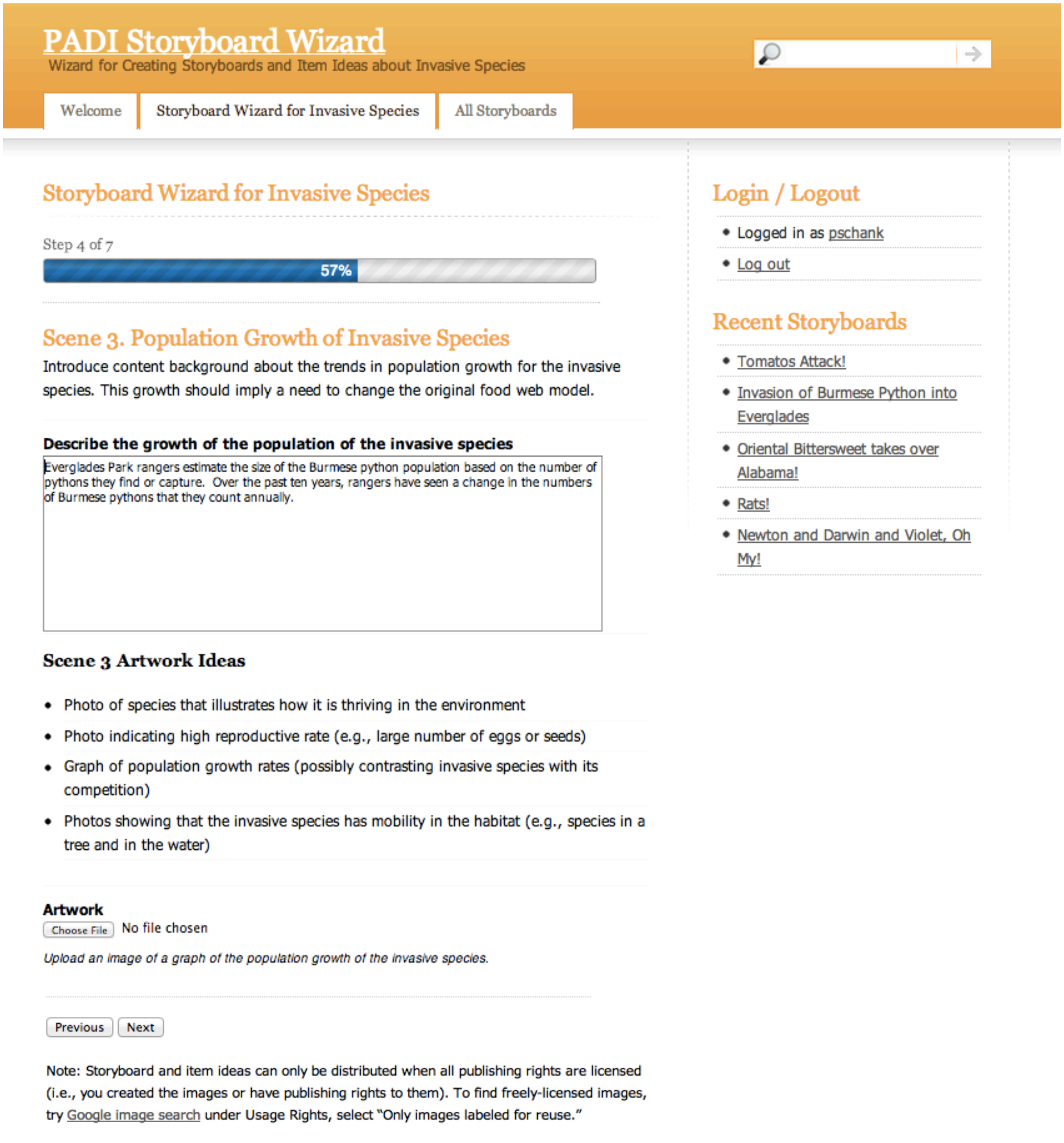


Figure 7: Storyboard Wizard Step 5: Describe where the invasive species enters the food web.

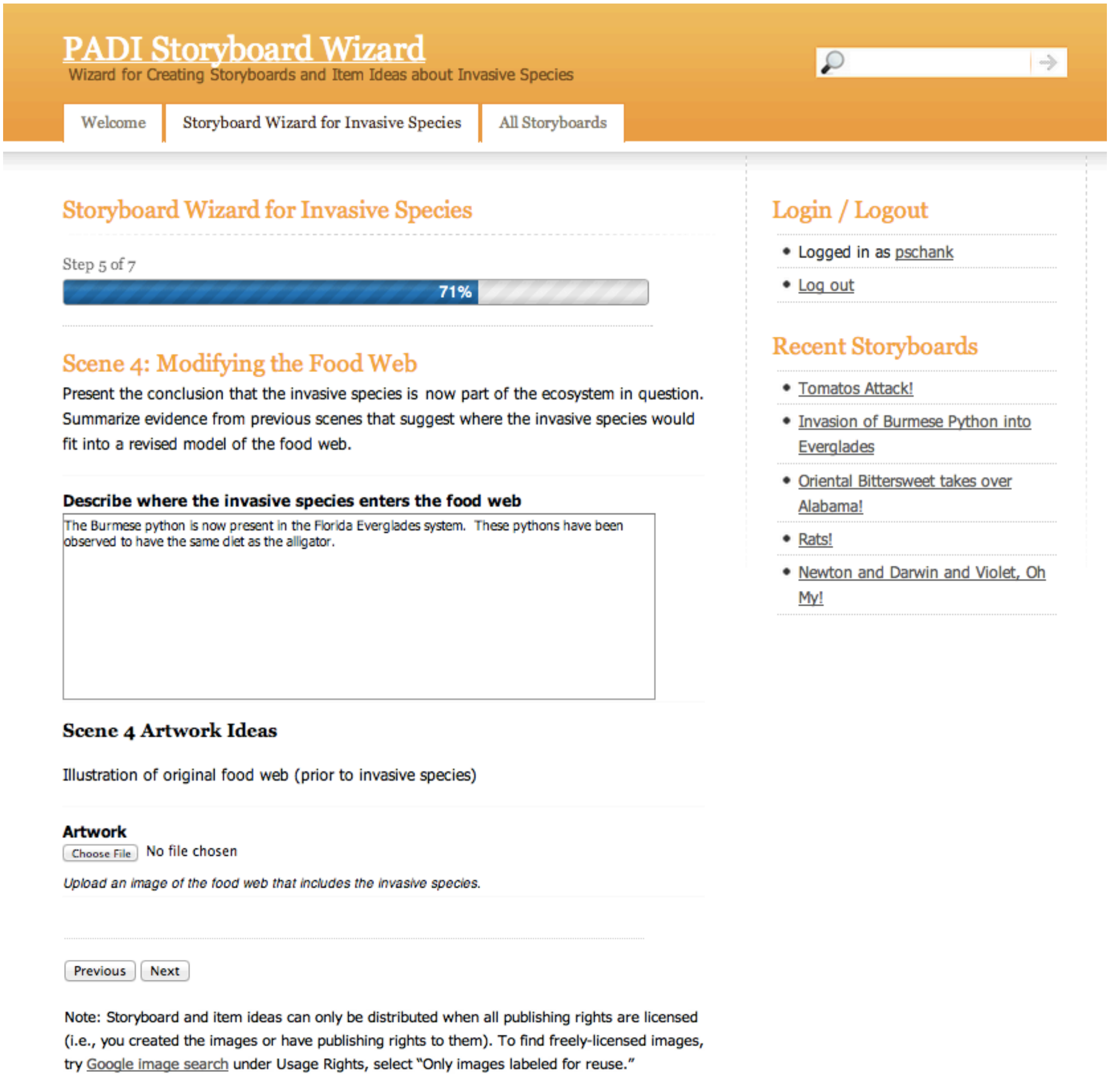


Figure 8: Storyboard Wizard Step 6: Describe the harmful effects of the growth of the invasive species.

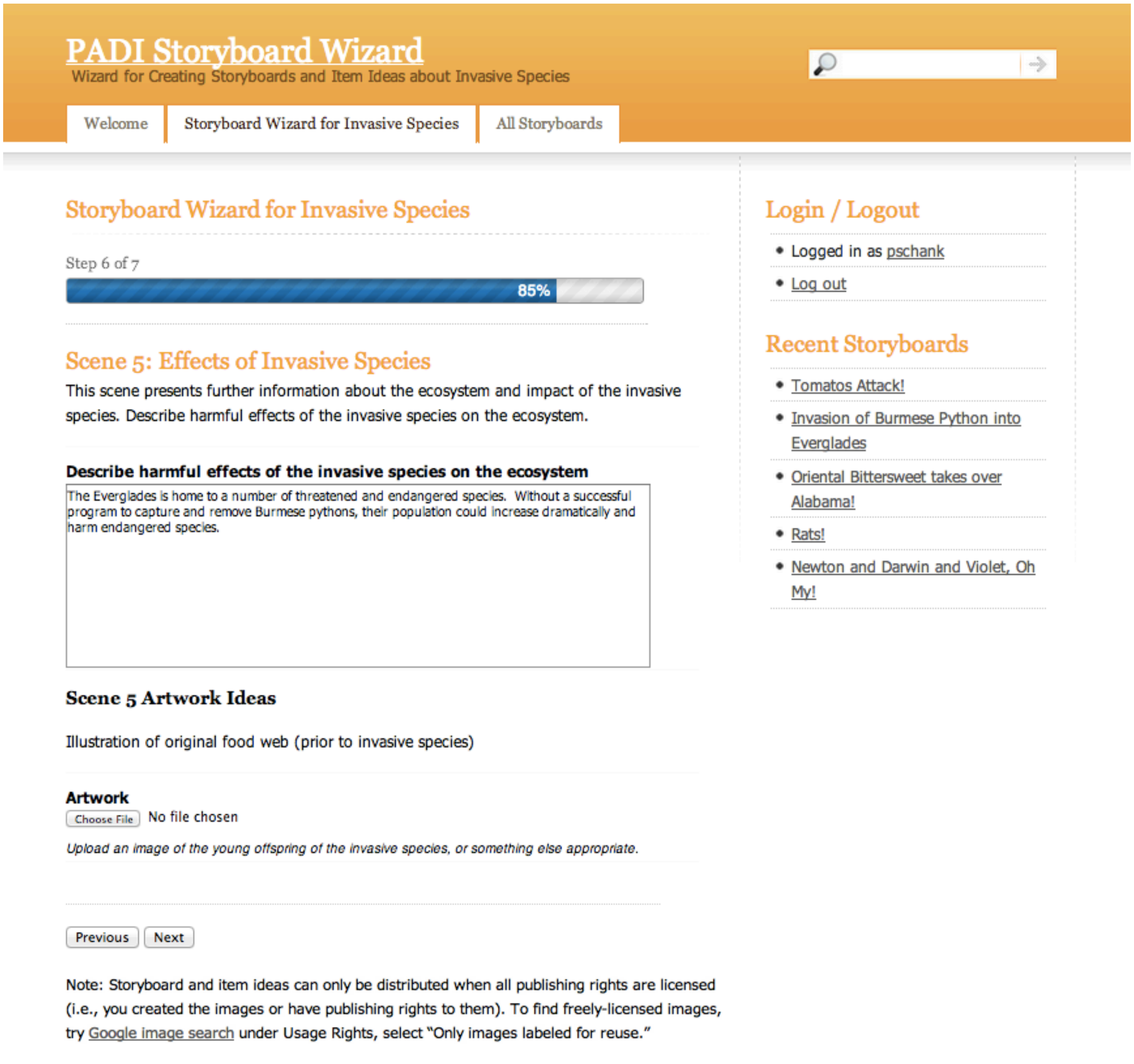


Figure 9: Example storyboard after completing the wizard (first two scenes).

PADI Storyboard Wizard

Wizard for Creating Storyboards and Item Ideas about Invasive Species

Welcome | Storyboard Wizard for Invasive Species | All Storyboards

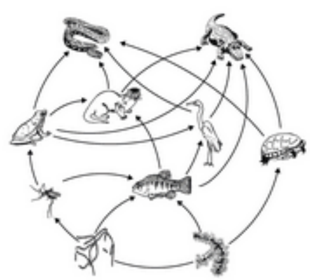
Invasion of Burmese Python into Everglades

mfeng September 14, 2011 | Print this storyboard | Show Outline | Export to XML

- Storyboard Type: Top Predator
- Invader Species: Burmese Python
- Publishing Rights: All uploaded images can be published (I created the images or have publishing rights to them)

Scene 1: Original Food Web

Everglades National Park is a large, warm, wetlands habitat in Florida. The food web for the Everglades includes many types of animals such as fish, birds, reptiles, insects, and mammals. Some Everglades animals are threatened species.



Items

+ Add

Scene 2: Introduction of the Invasive Species

In the past decade a new snake, the Burmese python, has invaded the Everglades ecosystem. Starting in the 1990s, some pet owners released pythons into the Everglades. This large snake thrives in warm and wet habitats and is the top predator in its native southeast Asia habitat. Scientists in Florida have recorded the content of pythons' stomachs to include many types of mammals and birds. The diet of python is illustrated in the artwork below.

Stomach Contents of Burmese Pythons in Florida

Mammals	Birds	Reptiles
• Rabbit • Hispid Cotton Rat • Cotton Mouse • Gray Squirrel • Fox Squirrel • Raccoon • Old World Rats • Virginia Opossum • Bobcat • Round-tailed Muskrat • Rice Rat • White-tailed Deer	• Pied-build Grebe • Limpkin • White Ibis • American Coot • House Wren	• Spotted Turtle • Eastern Indian Snake • Barbour's Map Turtle

Items

+ Add

Login / Logout

- Logged in as pschank
- Log out

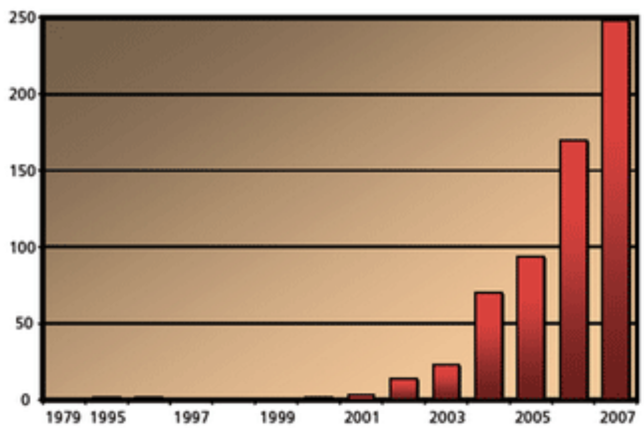
Recent Storyboards

- Invasion of Burmese Python into Everglades National Park
- Tomatos Attack!
- Invasion of Burmese Python into Everglades
- Oriental Bittersweet takes over Alabama!
- Rats!

Figure 10: Items to assess student understanding can be added from each completed scene.

Scene 3: Population Growth of the Invasive Species

"Everglades Park rangers estimate the size of the Burmese python population based on the number of pythons they find or capture. Over the past ten years, rangers have seen a change in the numbers of Burmese pythons that they count annually."



Yearly numbers of recovered pythons in Everglades National Park and surrounding areas. Data provided by Skip Snow, ENP

Items

- Why might the Burmese python population be growing?

+ Add

Figure 11: Items can be created from scratch or by adapting an existing item.

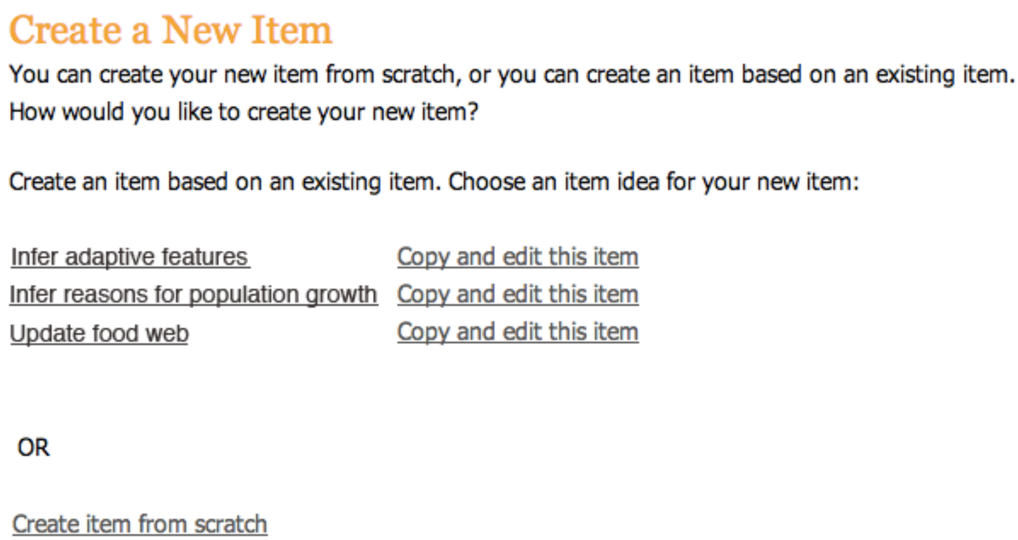


Figure 12: Creating a new item.

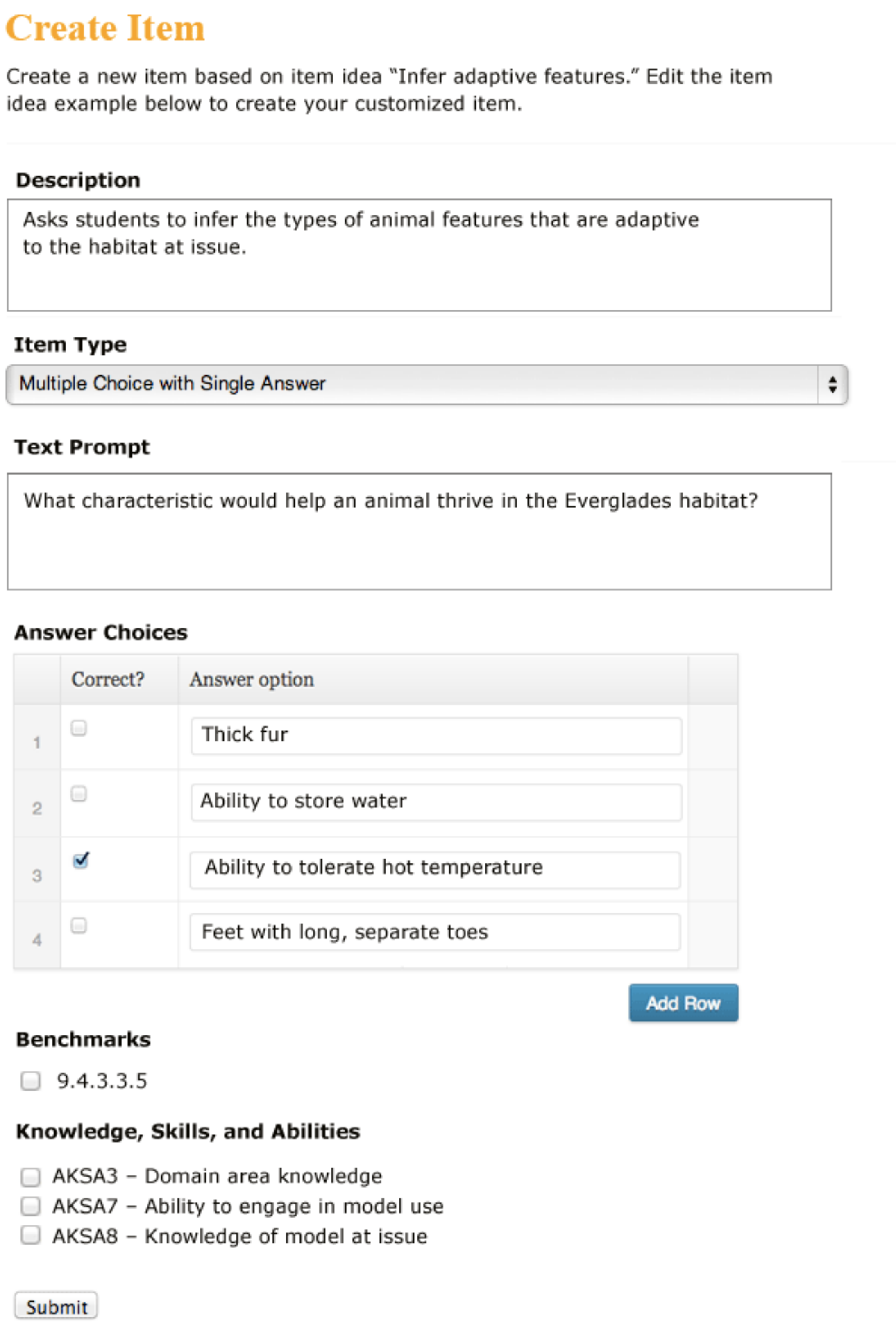


Figure 13: Example item customized from an item idea.

Infer Adaptive Features

Author: system | Created on 22-09-2011 | Edit this item idea | Delete this item

Description

Asks students to infer the types of animal features that are adaptive to the habitat at issue.

Item Type

Multiple Choice with Single Answer

Text Prompt

What characteristic would help an animal thrive in the Everglades habitat?

Answer Choices

Correct?	Answer
	Thick fur
	Ability to store water
correct	Ability to tolerate hot temperature
	Feet with long, separate toes

Benchmarks

9.4.3.3.5

Knowledge, Skills and Ability

AKSA3 – Domain area knowledge
AKSA7 – Ability to engage in model use
AKSA8 – Knowledge of model at issue

Figure 14: Example storyboard after adding an item for each scene (all scenes shown).

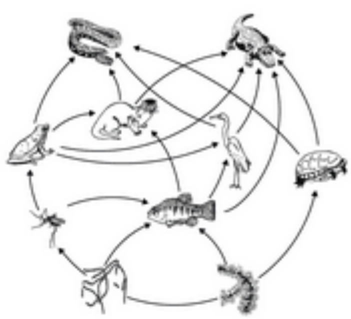
Invasion of Burmese Python into Everglades

mfeng September 14, 2011 | Print this storyboard | Show Outline | Export to XML

- Storyboard Type: Top Predator
- Invader Species: Burmese Python
- Publishing Rights: All uploaded images can be published (I created the images or have publishing rights to them)

Scene 1: Original Food Web

Everglades National Park is a large, warm, wetlands habitat in Florida. The food web for the Everglades includes many types of animals such as fish, birds, reptiles, insects, and mammals. Some Everglades animals are threatened species.



Items

- What characteristic would help an animal thrive in the Everglades habitat?

+ Add

Scene 2: Introduction of the Invasive Species

In the past decade a new snake, the Burmese python, has invaded the Everglades ecosystem. Starting in the 1990s, some pet owners released pythons into the Everglades. This large snake thrives in warm and wet habitats and is the top predator in its native southeast Asia habitat. Scientists in Florida have recorded the content of pythons' stomachs to include many types of mammals and birds. The diet of python is illustrated in the artwork below.

Stomach Contents of Burmese Pythons in Florida		
Mammals	**Birds**	**Reptiles**
• Rabbit • Hispid Cotton Rat • Cotton Mouse • Gray Squirrel • Fox Squirrel • Raccoon • Old World Rats • Virginia Opossum • Bobcat • Round-tailed Muskrat • Rice Rat • White-tailed Deer	• Pied-build Grebe • Limpkin • White Ibis • American Coot • House Wren	• Spotted Turtle • Eastern Indian Snake • Barbour's Map Turtle

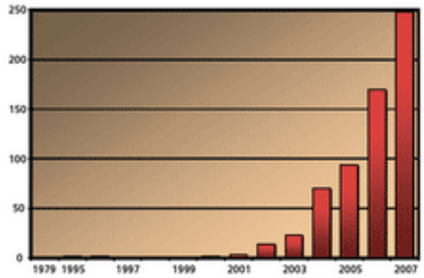
Items

- Based on the Burmese python's diet, how could a large number of Burmese pythons affect the Florida Everglades food web?

+ Add

Scene 3: Population Growth of the Invasive Species

Everglades Park rangers estimate the size of the Burmese python population based on the number of pythons they find or capture. Over the past ten years, rangers have seen a change in the numbers of Burmese pythons that they count annually.



Yearly numbers of recovered pythons in Everglades National Park and surrounding areas. Data provided by Skip Snow, ENP

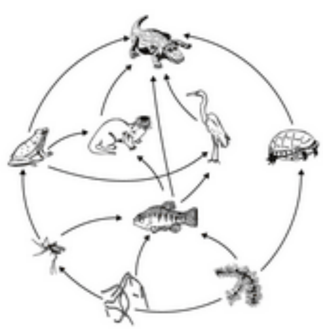
Items

- Why might the Burmese python population be growing?

+ Add

Scene 4: Modifying the Food Web

The Burmese python is now present in the Florida Everglades system. These pythons have been observed to have the same diet as the alligator.



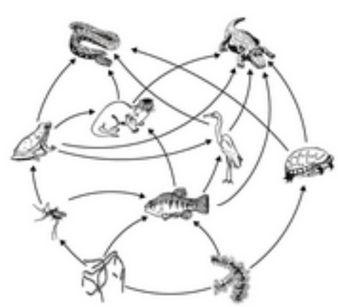
Items

- How should the original Everglades food web be updated?

+ Add

Scene 5: Effects of the Invasive Species

The Everglades is home to a number of threatened and endangered species. Without a successful program to capture and remove Burmese pythons, their population could increase dramatically and harm endangered species.

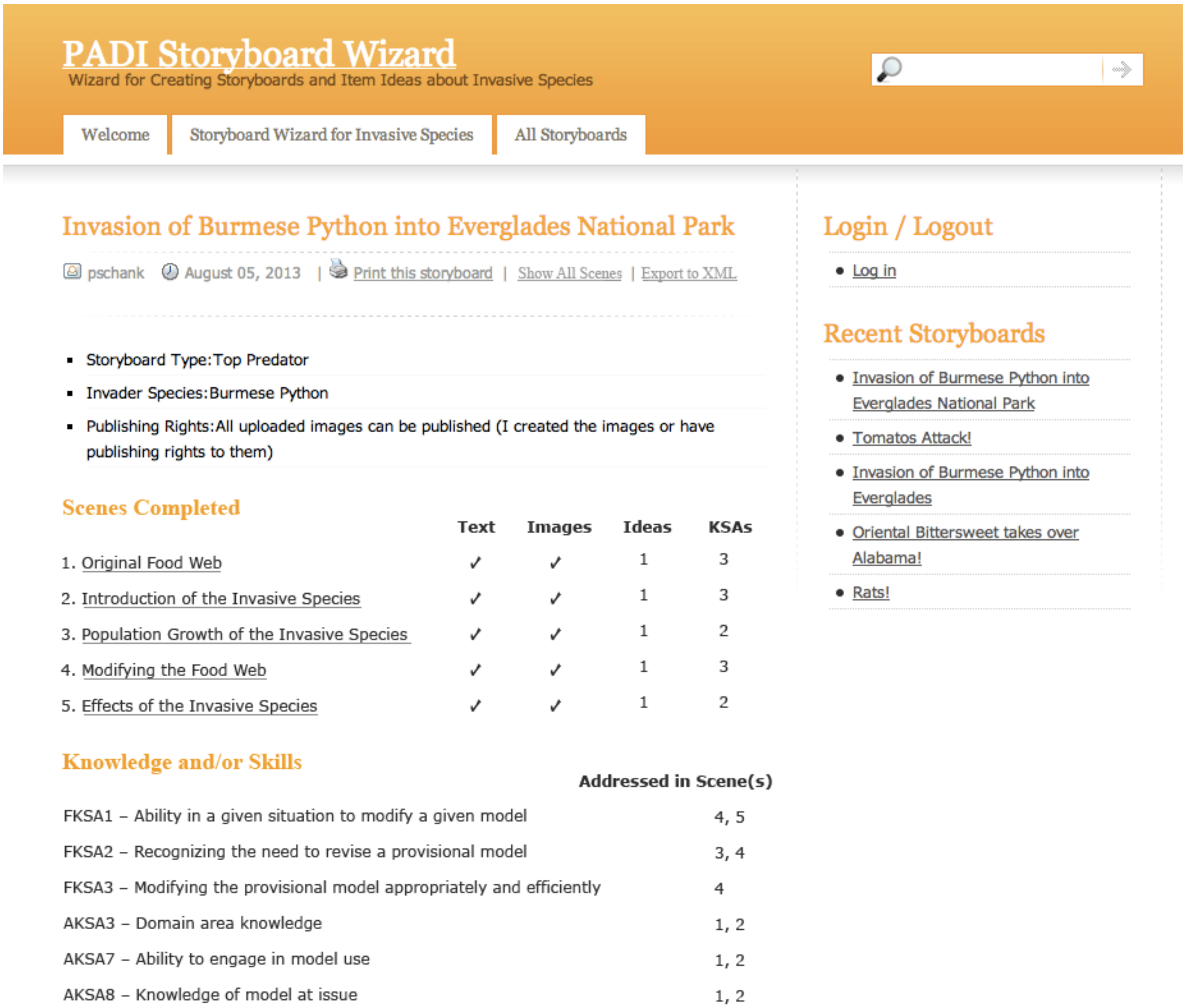


Items

- What are some potential impacts of an increased Burmese python populationl in the Everglades?

+ Add

Figure 15. Sample outline of all scenes and KSAs addressed.



PADI Storyboard Wizard

Wizard for Creating Storyboards and Item Ideas about Invasive Species

Welcome | Storyboard Wizard for Invasive Species | All Storyboards

Invasion of Burmese Python into Everglades National Park

pschank | August 05, 2013 | Print this storyboard | Show All Scenes | Export to XML

- Storyboard Type:Top Predator
- Invader Species:Burmese Python
- Publishing Rights:All uploaded images can be published (I created the images or have publishing rights to them)

Scenes Completed

	Text	Images	Ideas	KSAs
1. Original Food Web	✓	✓	1	3
2. Introduction of the Invasive Species	✓	✓	1	3
3. Population Growth of the Invasive Species	✓	✓	1	2
4. Modifying the Food Web	✓	✓	1	3
5. Effects of the Invasive Species	✓	✓	1	2

Knowledge and/or Skills

	Addressed in Scene(s)
FKSA1 – Ability in a given situation to modify a given model	4, 5
FKSA2 – Recognizing the need to revise a provisional model	3, 4
FKSA3 – Modifying the provisional model appropriately and efficiently	4
AKSA3 – Domain area knowledge	1, 2
AKSA7 – Ability to engage in model use	1, 2
AKSA8 – Knowledge of model at issue	1, 2

Login / Logout

- Log in

Recent Storyboards

- Invasion of Burmese Python into Everglades National Park
- Tomatos Attack!
- Invasion of Burmese Python into Everglades
- Oriental Bittersweet takes over Alabama!
- Rats!

5.0 Discussion

5.1 The Psychology of Task Design

Section 3 noted some initial findings on the psychology of item writing (Fulkerson, Nichols, & Mittelholtz, 2010; Nichols & Fulkerson, 2010). Item writing, and by extension the authoring of the more complex assessments such as the scenario-based tasks in the Minnesota Comprehensive Assessments, is an exercise in particular kind of problem solving, design under constraints (Akin, 1986; Katz, 1994; Simon, 2001).

There are constraints of several distinct kinds. No matter what kind of task is at issue, the substance of the task must be accurate, it must be suited to the intended examinees, it must fit the style and presentation requirements, and it must address the targeted level of knowledge. In a more complex task, there are issues of comprehensibility, flow among multiple aspects of a task, and coverage of multiple objectives. And there is always the need to embody a coherent assessment argument, satisfying the basic Messick structure of what kinds of things we must see students do in what kinds of situations, but also minimizing construct irrelevant demands.

Expert test developers solve these design challenges iteratively; from initial ideas that seem to hold promise, they successively revise provisional designs to meet more and more of the constraints. The more complex the intended task, the more constraints in number, in kind, and in interaction. The undertaking can be challenging to an expert, and overwhelming to novices.

Writers with experience in a domain, such as teaching science in the targeted grade levels, can draw on initial ideas from past personal experience with classroom exercises, curricular materials, interesting articles, and current news in science, all as potential specific contexts for an assessment task. Writers with expertise in assessment can draw on schemas and many examples of tasks that build in pieces of solutions to the multiple-constraint problem, which they can adapt to the context and the item-writing assignment at hand.

5.2 The Role of Wizards

At one end of a continuum, task developers are left totally to their own devices to solve these complex design-under-constraint problems. At the other end are fully-packaged solutions: procedures that can create tasks fully automatically, such as Bormuth's (1970) transformational grammar algorithm for generating test items from texts, Hively, Patterson, and Page's (1968) "item forms," and Embretson's (1998) cognitive design system. The generation of tasks for the British Army Recruitment Battery is in fact fully automated, and creates a unique parallel form of the test for every examinee, on the fly (Irvine, in press). In these instances, the design-under-constraint problem has been solved in its entirely remotely, by a team of experts. For at least a defined class of tasks, the user can obtain tasks with no knowledge at all—and, correspondingly, no opportunity to adapt them to local conditions.

Most work at large testing agencies is somewhere between these extremes, with familiar task schemas to help structure work and more experienced colleagues to offer advice. Most teacher-generated tasks are adaptations of instances of previously-used tasks by oneself or colleagues, or exemplars from professional publications or support materials. The former is particularly helpful when the intended tasks are similar to familiar ones, while the latter is more useful when new kinds of tasks with additional or unfamiliar constraints have been introduced into the problem.

Wizards offer a more structured point in the middle of the continuum. A wizard builds in some of the knowledge and some support to meet certain kinds of constraints, but leaves other choices up to the user. The complex multi-constraint design problem has been simplified somewhat, in that partial solutions to some sets of constraints are provided, and support is offered for choices to meet additional constraints. This current project's interviews with users of design patterns found that even expert developers experienced value from design patterns that activated memories of useful experiences for building out the specifics of tasks. Wizards are a kind of formalization of the useful-schema and advice-from-a-colleague kinds of support for test development.

The goal is to have the wizard allow flexibility for what developers are good at, while at the same time supporting the difficult aspects they are not necessarily good at (e.g., structuring a complex assessment argument in interactive or multiple steps tasks, and instantiating the results of the design processes in the data structures needed in the assessment system of interest).

5.3 What Should Be Supported?

Constraints of several kinds need to be addressed in the design of a complex assessment task. Which ones should a given wizard support, and how scripted should the support be? The answer depends on the anticipated users and what job they need to do.

Wizard designers must determine the level of specificity and support they want to build into a wizard: There is a tradeoff between, on the one hand, specificity and heavy scaffolding, and on the other, breadth of use and less scaffolding support. The more specific the class of tasks they want to support, the easier they can make the wizard for the user. The more flexible and wide ranging the wizard, the more responsibility for content knowledge and assessment design knowledge will be required on the part of the user.

The current wizard focuses on a particular class of investigation, namely model revision as a practice, in the content area of ecology, with an invasive species story line. There is strong support for this particular story line, and within in it, very strong support for building stimulus materials and items for assessing important science objectives. Model revision in particular is a hard-to-assess process skill. The PADI Model Revision design pattern (Mislevy, Riconscente, Rutstein, 2009) indicates the kinds of characteristic features that must be in a task to assess model revision. It notes features that can be varied; concomitant additional knowledge and skill that can (at the designer's discretion) be additionally required; features of tasks that can be varied to stress different aspects of knowledge; work products and observable variables that can employed; and narrative structures that work well for assessing this process skill – all to the end of helping the designer build a task that embodies coherent and valid assessment argument.

These are all assessment design considerations that evolved from cutting edge research on assessment design (ECD in particular) that are not generally familiar to task authors who are not assessment experts—and even many who are.

The wizard has built into it a scenario based task that addresses these assessment design considerations at a level of generality that is a step above a specific task.

What is not build in, and what remains for the designer to create as the wizard walks her through the design process, is the detailing of a particular invasive species, and items that are specific to the food web in the ecosystem, and the relationships among its denizens and ecological features. These choices require two critical areas of local knowledge the wizard cannot in and of itself supply, and the user must provide.

The first is science content knowledge. In choosing the ecosystem and invasive species, the user must either know or have external support for its particulars, and be able to tailor the item ideas the wizard supplies to these particulars.

The second is knowledge of the intended testing population. How complex should the food web be? Should the ecosystem be one that is familiar to the students, so they can rely on personal knowledge, or should they be unfamiliar so that they need to lean more heavily on science principles? Should items about the environment and about the animals in question be more technical? Should they build on recent work in class? Should they be in choice or open-ended formats?

This wizard, then, is tuned to the user who knows the science, knows the students, and knows the testing purpose, but is probably not an expert test developer so that the less-familiar assessment argumentation constraints are the ones that are supported. An opposite kind of user, who knows assessment design but not the underlying science or what students at a given level think about, would still have a hard time creating an invasive species model-revision scenario-based task with the wizard—although there is enough content support to focus the research this user would need to do.

5.4 The Role of ECD structures

Although the idea of a wizard to support assessment design is not new (e.g., ETS's 2000 *Assessment Wizard*), what is innovative about the wizard described in this report is its grounding on evidence-centered design principles and structures. Certain efficiencies in wizard design, validation, and moving to implementation are achieved with this foundation.

- Regarding wizard design: The wizard is grounded on design choices built into the attributes of a PADI design pattern. Both the choices for the wizard designer for creating the basic task structure and the choices to be provided to the user for detailing out the task are based on the essential design space laid out in the form of the assessment design argument, and built into the form of the PADI design pattern. This is a theoretically grounded starting point for designing any wizard, in contrast to wizards that are based on the surface features, say of task types or rubrics. Again, this doesn't mean that designing a wizard will be easy; it just means that one knows how to think about it, from a theory-based model of assessment, and practical tools that have been developed in PADI to support assessment design work.

- Regarding validity: One strand of validity argumentation is called construct representation (Embretson, 1983): It concerns the grounding and the design rationale for the elements and procedures in an assessment task. In other words, just why do we think observing these particular aspects of what a student says or does, in a situation with these particular features and affordances, ought to give us evidence about the targeted knowledge and skills? The underlying PADI design pattern supports such a case for tasks produced using a wizard based on it. *The design pattern builds on research and experience for the capability of interest, and lays out a space of design choices that are both consistent with this research and organized to form a coherent assessment argument.* In other words, key constraints have been identified and a supporting space has been built for the user to make final choices to produce a specific task. Of course simply using a wizard does not guarantee a good or a valid task. It is still quite possible to choose poorly in a well designed and well supported design space. What we can do from a distance, as wizard designers, is increase the odds.

- Regarding implementation: While the preceding points emphasized the role of PADI design patterns, efficiencies for implementation are gained more through PADI template objects and task objects, and comparable structures (e.g., Luecht, 2013; Mislevy, Behrens, et al., 2010). The key is that built into the wizard is the capability to pre-arrange the integration of the task schema the wizard is built around and the user-provided information needed to complete it in the form of the objects that the intended delivery system will use. This can be very specialized knowledge that neither content experts nor test developers typically have. Usually in technology-delivered assessment systems there is a 'handoff' from developers who create the content of a task to the technical team that implements it. This work can be built into the wizard structure, once in a more general form, by the group of experts who designs the wizard. Examples of wizards that do this are the PADI Mystery Powders wizard (Hamel, Mislevy, & Winters, 2008) and the Cisco Networking Academy design interface for troubleshooting tasks in the Packet Tracer simulation tool (Frezzo, Behrens, & Mislevy, 2009).

5.5 Final comment

Exciting advances in assessment are better understandings of learning in content domains, broader ranges of ways we can conceive to assess this learning, and digital environments that enable us to implement richer and more interactive assessments. A remaining bottleneck is figuring out just how to do this in efficient and valid ways. Recent advances in the science of assessment theory provide us with conceptual tools for doing this better (National Research Council, 2002), and advances in "assessment engineering" help us take advantage of technology to put this knowledge into practice (Mislevy, Bejar, Bennett, Haertel, & Winters, 2010). Building across these advances, capitalizing on general research and allowing local customization, assessment design wizards can be particularly handy tools our toolkit.

References

Akin, O. (1986). Psychology of architectural design. London: Pion.

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

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. Available online at http://padi.sri.com/downloads/TR5_IDFramework.pdf

Bormuth, J.R. (1970). *On the theory of achievement test items.* Chicago: University of Chicago Press.

Educational Testing Service (2000). *Assessment Wizard: A Professional Development Tool for Designing and Sharing Student Assessments* [Computer software]. Princeton NJ: Educational Testing Service.

Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

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/

Fulkerson, D., Nichols, P., & Mittelholtz, D. (2010). What item writers think when writing items: Towards a theory of item writing expertise. Presented at the American Educational Research Association Annual Meeting, Denver, Colorado, May 3, 2010. Available online at http://www.pearsonassessments.com/NR/rdonlyres/E86B7038-46AA-4CE7-9BE2-520FDD3ABB46/0/2_AERA2010TheoryItemWritingPaper_42310.pdf

Hamel, L., Mislevy, R., & Winters, F. (2008). *Design Rationale for an Assessment Task Authoring System: A Wizard for Creating "Mystery Inquiry" Assessment Tasks* (PADI Technical Report 19). Menlo Park, CA: SRI International.

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275-290.

Irvine, S.H. (in press). *Tests for recruitment across cultures: A tactical psychometric handbook.* IOS Press: Amsterdam, The Netherlands.

Katz, I.R. (1994). Coping with the complexity of design: Avoiding conflicts and prioritizing constraints. In A. Ram, N. Nersessian, & M. Recker (Eds.), *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society* (485-489). Mahwah, NJ: Erlbaum.

Liu, M., & Haertel, G. (2011). Design patterns: A tool to support assessment task authoring (*Large-Scale Assessment Technical Report 11*). Menlo Park, CA: SRI International.

Luecht, R.M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing*

Technology, 14. Available online at http://www.testpublishers.org/assets/documents/test%20specifications%20jatt%20special%20issue%2013.pdf

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

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* (Fourth Edition) (pp. 257-305). Phoenix, AZ: Greenwood.

Mislevy, R.J., Almond, R.G., & Lukas, J. (2004). A brief introduction to evidence-centered design. *CSE Technical Report 632.* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K., & Winters, F.I. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment, 8*(2). http://escholarship.bc.edu/jtla/vol8/2

Mislevy, R.J., Bejar, I.I., Bennett, R.E., Haertel, G.D., & Winters, F.I. (2010). Technology supports for assessment design. In B. McGaw, E. Baker, & P. Peterson, (Eds.), *International Encyclopedia of Education,* 3rd Edition, Volume 8 (pp. 56-65). Amsterdam: Elsevier.

Mislevy, R.J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25,* 6-20.

Mislevy, R., Hamel, L., Fried, R., G., Gaffney, T., Haertel, G., Hafter, A., Murphy, R., Quellmalz, E., Rosenquist, A., Schank, P., Draney, K., Kennedy, C., Long, K., Wilson, M., Chudowsky, N., Morrison, A., Pena, P., Songer, N., Wenk, A. (2003). Design patterns for assessing science inquiry (PADI Technical Report 1). Menlo Park, CA: SRI International. Also presented at American Education Research Association (AERA) in April, 2003. Available online at http://padi.sri.com/downloads/TR5_IDFramework.pdf

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., Riconscente, M.M., & Rutstein, D.W. (2009). Design patterns for assessing model-based reasoning *(PADI-Large Systems Technical Report 6).* Menlo Park, CA: SRI International.

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

National Research Council (2002). Advances in the sciences of thinking and learning. In J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), *Knowing what students know* (pp. 59-110). Washington, D.C.: National Academy Press.

Nichols, P., & Fulkerson, D. (2010). Informing design patterns using research on item writing expertise. (*Large-Scale Assessment Technical Report 9*). Menlo Park, CA: SRI International. Available online at http://ecd.sri.com/downloads/ECD_TR9_Item_Writing_Expertise.pdf

Riconscente, M., Mislevy, R., Hamel, L., & PADI Research Group (2005). An introduction to PADI task templates (PADI Technical Report 3). Menlo Park, CA: SRI International. Available online at http://padi.sri.com/downloads/TR5_IDFramework.pdf

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.

Simon, H.A. (2001). *The sciences of the artificial* (4th ed.). Cambridge, MA: MIT Press.

Sponsor
The National Science Foundation, Grant DRL - 0733172

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

Subgrantees
University of Maryland
Pearson