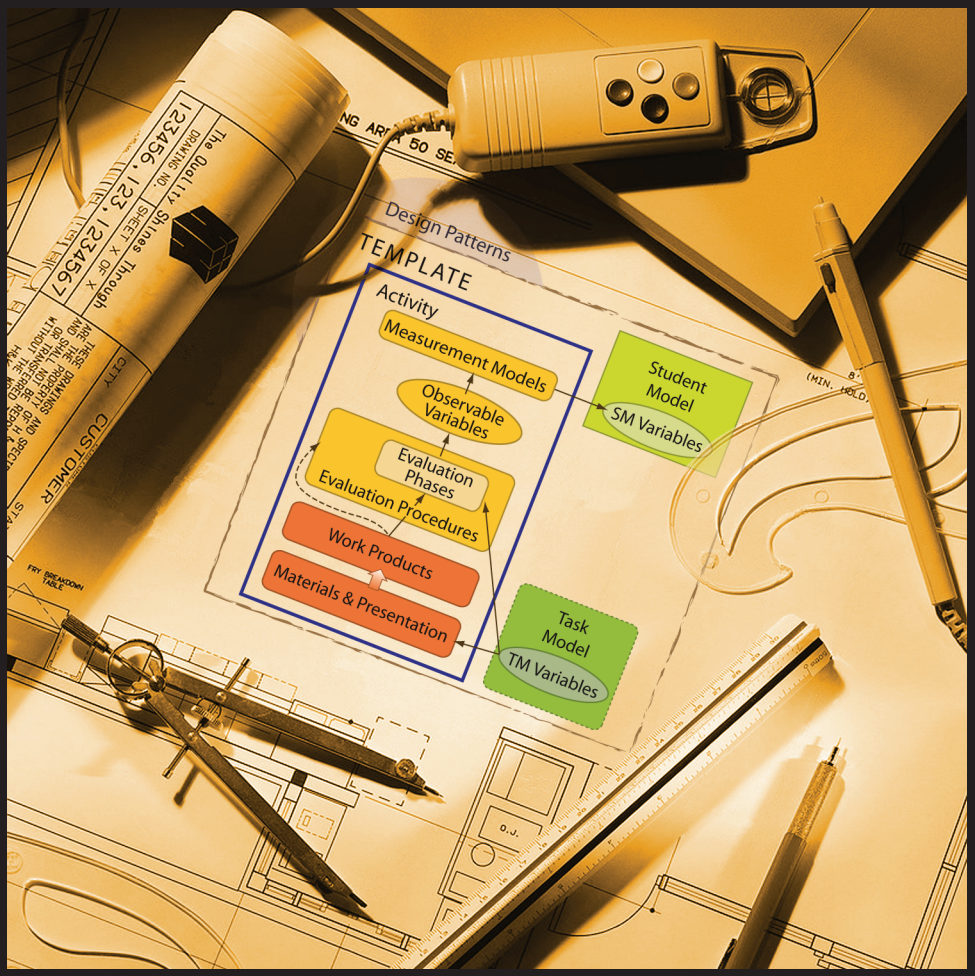


PADI Technical Report 15 | September 2006

Mystery Powders: An Application of the PADI Design System Using the Four-Process Delivery System

PADI | Principled Assessment Designs for Inquiry

Graham Seibert, University of Maryland
Lawrence Hamel, CodeGuild, Inc.
Kathleen C. Haynie, Kathleen Haynie Consulting
Robert J. Mislevy, University of Maryland
Han Bao, 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://padi.sri.com

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

PRINCIPLED ASSESSMENT DESIGNS FOR INQUIRY
TECHNICAL REPORT 15

Mystery Powders: An Application of the PADI Design System Using the Four-Process Delivery System

Prepared by:
Graham Seibert, University of Maryland
Larry Hamel, CodeGuild, Inc.
Kathleen Haynie, Kathleen Haynie Consulting
Robert Mislevy, University of Maryland
Han Bao, University of Maryland

Acknowledgments
This material is based on work supported by the National Science Foundation under grant REC-0129331 (PADI Implementation Grant). We are grateful for contributions by Geneva Haertel, Cathleen Kennedy, and Futoshi Yumoto.
Disclaimer
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

CONTENTS

FIGURES

TABLES

ABSTRACT

This report illustrates how the general principles and structural components of the PADI framework were applied to a Mystery Powders assessment demonstration project that was computer-based. The report begins with brief overviews of the Mystery Powders assessment and the PADI design framework. We then describe the classic hands-on Mystery Powders chemistry experiment in more depth, our team's objectives in implementing a computer-based version of the task using the PADI design system and four-process delivery architecture (Almond, Steinberg, & Mislevy, 2002), the differences between the original and the computer-delivered assessments, and the technical challenges we faced. The PADI synopsis describes the design layers in the PADI framework, briefly introducing key design structures in each layer. These layers are *domain analysis*, *domain modeling*, *conceptual assessment framework*, *assessment implementation*, and *assessment delivery*. For purposes of our demonstration, a computerized adaptive test (Wainer, 2000) was presented that enabled an examinee to work through a series of Mystery Powders tasks. Through the logic of Lord's (1971, 1980) "flexilevel" adaptive testing scheme, a harder task was presented after a successful solution and an easier task was presented after an unsuccessful solution. The body of this report addresses the design and technical issues that arose in the implementation of a computer-based interactive assessment, carried out with the PADI design system and cast in the four-process delivery architecture.

1.0 Introduction and Overview

Principled Assessment Designs for Inquiry (PADI) is a project supported by the National Science Foundation to improve the assessment of science inquiry.[1] The PADI project has developed a design framework for assessment tasks, based on the evidence-centered design (ECD) framework introduced by Mislevy, Steinberg, and Almond (2002). PADI was developed as a system for designing blueprints for assessment tasks, with a particular eye toward science inquiry tasks—tasks that stress scientific concepts, problem solving, building models, using models, and cycles of investigation. The PADI framework guides an assessment developer's work. Its design structures, expressed in terms of extensible object models[2], guide the developer to consider all the relevant factors in assessment development and to take advantage of the commonalities between the assessment being developed, existing assessments, and prospective assessments (that can share conceptual and operational elements). PADI is meant to integrate the processes of assessment design, authoring, delivery, and scoring to ensure that critical considerations (e.g., consistency, usability, validity) inform the process from its inception. This report illustrates how the general principles and structural components of the PADI framework are applied in the Mystery Powders assessment demonstration project.

We will begin with an overview of the PADI design framework. The PADI synopsis will describe the design layers in the PADI framework, briefly introducing key design structures in each layer. These layers are *domain analysis*, *domain modeling*, *conceptual assessment framework*, *assessment implementation*, and *assessment delivery*. We then will offer a synopsis of Mystery Powders by describing the classic hands-on Mystery Powders chemistry experiment, our team's objectives in implementing a computer-based version of the task using the PADI design system and four-process delivery architecture (Almond, Steinberg, & Mislevy, 2002), the differences between the original and the computer-delivered assessments, and the technical challenges faced in delivering the computer-based version of Mystery Powders. Because this version of the Mystery Powders assessment is delivered via the QTI[3] protocol, we call it MP-QTI. For purposes of our demonstration, a computerized adaptive test (Wainer, 2000) is presented. With computerized adaptive testing (CAT), an examinee works through a series of Mystery Powders tasks. Through the logic of Lord's (1971, 1980) "flexilevel" adaptive testing scheme, a harder task is presented after a successful solution and an easier task is presented after an unsuccessful solution. The body of this report addresses the design and technical issues that arise in the implementation of a computer-based interactive assessment, designed in the PADI design system, and cast in the four-process delivery architecture.[4]

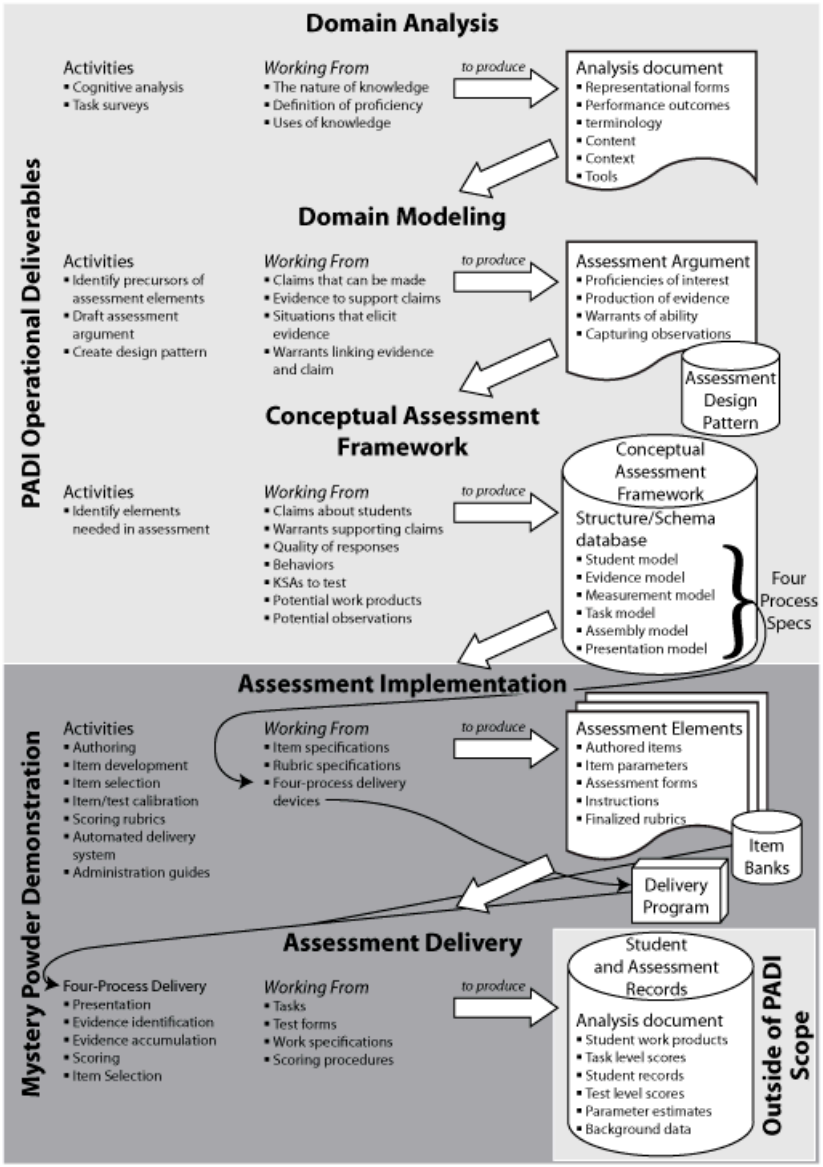
---

[1] PADI is a collaboration among researchers and developers at SRI International, the University of Maryland, the University of California-Berkeley, the University of Michigan, and Lawrence Hall of Science.
[2] The reader is referred to Rumbaugh, Jacobson, & Booch (1998) for an overview of an object modeling approach to software design, and the application of these ideas to modeling business or other systems.
[3] Question and Test Interoperability (QTI) standards specified by IMS Global Learning Consortium, Inc. (2000).
[4] Note that the primary purpose of MP-QTI is to demonstrate the use of PADI design object model and the four-process delivery system, rather than cognitive analysis of mystery powders problem-solving, instructional or decision-making usage, optimal scoring, or any other of many issues that could be explored further. We lean instead on previous research to ground plausible design decisions for illustrating use of the design machinery.

1.1 Introduction to the PADI Design Framework and Design Layers

To sort out the many elements of the PADI design framework and explain their roles in the Mystery Powders example, it is useful to organize the design and delivery of an assessment in terms of layers (Mislevy & Riconscente, 2005). Within a complex process such as assessment design, layers identify subsystems (the individual components of which are better handled at the subsystem level). Each layer in Figure 1 identifies a conceptual, structural, or operational level of activities and products in the assessment enterprise. Although the types of knowledge representations within each layer are fairly self-contained, layers inform and are informed by other layers. The layers are defined as *domain analysis*, *domain modeling*, *conceptual assessment framework*, *assessment implementation*, and *assessment delivery*.

Figure 1. PADI Evidence-Centered Design: Stages and Work Products Overlaid with Four Process



The *domain analysis* layer identifies valued knowledge in the domain of interest (e.g., what examinees need to know), situations in which people would use this knowledge as they interact with the environment and other people, and how people use their knowledge (including rich descriptions of tasks in which the knowledge is utilized). Examples of *domain analysis* activities include literature reviews, job analyses, and cognitive task analyses. The particular sources of information and forms of knowledge and situations vary widely from one domain to another. Because PADI is intended to be used across domains, PADI does not provide specific tools or procedures for *domain analysis* within the realm of scientific inquiry. The structure imposed by PADI *design patterns* and *templat*es, however, can usefully shape the way *domain analysis* is carried out—specifically, in ways and using tools that lend themselves to thinking about the pertinent assessment questions (e.g., Table 1). For example, Shute, Torreano, and Willis's (2000) automated knowledge elicitation and organization tool DNA (for Decompose, Network, Assess) could be tuned to provide input to structures in the *domain modeling* layer of ECD.

The *domain modeling* layer provides articulation between the wealth of information about the domain gathered in *domain analysis* and the technical, rather esoteric, elements of the Conceptual Assessment Framework. Among the representational forms that are used for *domain modeling* are Toulmin diagrams (Mislevy, 2003) and PADI *design patterns* (PADI, 2003). In the *domain modeling* layer, information needed for the three key elements of the assessment argument - the Student, Task, and Evidence Models - is organized in a narrative rather than technical form.

Four critical questions guide the evidence-centered approach to assessment design in the area of science inquiry (Baxter & Mislevy, 2005, p.2):

1. What does it mean to know and do inquiry?
2. What constitutes evidence of knowing?
3. How can that evidence be elicited from examinees?
4. What are the appropriate statistical techniques for making valid inferences about what examinees know from what examinees do?

Design patterns, implemented as Web-based forms, are used to structure information (in terms of assessment arguments) in any content area and from any psychological perspective (e.g., cognitive, sociocultural), providing a theoretical underpinning for design practices. Two *design pattern*s applicable to Mystery Powders are shown later in Figure 2 and Figure 5. These are typical of forms in the PADI design system. A *design pattern*, or other PADI form, consists of prescribed fields that are common to all elements (or entities) of its own type and links to other entities to which it is related, either superior,

subordinate, or in some way parallel[5]. The content of these *design patterns* will be discussed further in the *domain modeling* section of the paper. It suffices at this point to note that the first (Table 3) pertains to assessment arguments for a broad class of potential tasks, namely hypothetico-deductive problem-solving in finite domains[6]. The second (Figure 5), a special case of the first, pertains specifically to Mystery Powders tasks in a variety of modes (e.g., lab-based, simulated, talk-alouds). All PADI forms, particularly those utilized at the *conceptual assessment framework* layer, share this object-oriented structure. This structure provides a formal mechanism that:

- ensures that assessment designers do not overlook significant elements of information,
- allows one entity to use - that is to borrow by reference - attributes of another entity,
- documents the web of relationships among entities, and
- makes it easy to create new entities by a process of clone-and-modify.

The *conceptual assessment framework* specifies the components and technical machinery that embody the assessment argument (Mislevy & Riconscente, 2005; Mislevy, Steinberg, & Almond, 2002). If an analogy is made between an assessment and building project, the *design pattern* corresponds to the conceptual sketches and site model, while the CAF is analogous to the detailed blueprints (tasks and rubrics are the sticks and bricks). Table 1 shows the elements of the CAF and their relationship to critical questions. Working from qualitative descriptions of assessment argument components, the substantive, technical, and operational details that are required to provide an evidence-centered assessment design are specified in the CAF layer. This critical stage is supported in PADI through the use of objects (and accompanying web-based forms) referred to as *templates*. The assessment designer completes the *templates*, which, like *design patterns*, can be applied across content areas, assessment purposes, and psychological perspectives.

---

[5] The expression of the design objects in PADI is compatible with the IMS/QTI standards for electronic learning and assessment objects. IMS is the short name for the Global Learning Consortium, Inc., which was originally the Instructional Management Systems (IMS) project when it was formed in 1997. IMS is concerned with establishing interoperability for learning systems and learning content and the enterprise integration of these capabilities. PADI's IMS compatibility promotes the reuse of PADI assessment objects by different programs, content developers, and service providers (e.g., for assessment delivery processes such as presentation and task scoring).

[6] The hypothetico-deductive model is a theory about scientific method. According to the theory, scientific inquiry proceeds by formulating a hypothesis that is intended to explain an observed phenomenon, and from the hypothesis a sufficient number of explicit predictions of further phenomena are deduced that should be observable as a consequence of the hypothesis. Observations that run contrary to those predicted are taken as a conclusive falsification of the hypothesis, and observations that are in agreement with those predicted are taken as corroborating the hypothesis. It is then supposedly possible to compare the explanatory value of competing hypotheses by looking to see how well they are sustained by their predictions.

Table 1. Association of CAF Components to Baxter & Mislevy (2005) Questions

CAF component	Description	Relation to the Questions
Student Model	A model of the examinee Knowledge, Skills and Abilities (KSAs) to be addressed by the assessment, possibly multidimensional.	"What does it mean to know and do inquiry?" This could include a number of dimensions such as subject matter knowledge, designing and conducting experiments, using tools and techniques to gather data, developing descriptions, making connections between evidence and explanations, and communicating results.
Evidence Model	A model of observations that may be made about examinees, including the evaluative plan for converting examinee Work Products into numeric scores, and a Measurement Model for relating observations to examinee-model variables.	"What constitutes evidence of knowing?" How, for instance, would one deduce a examinee's understanding of modes of inheritance? Behavior in the field? Responses to a set of questions? And, how will it be measured? What are the appropriate statistical techniques for making valid inferences about what examinees know from what examinees do?
Task Model	A model of the tasks that can be used to elicit evidence of examinee KSAs, identifying Work Products, materials and specifications, and fixed and Variable Features associated with a task.	"How can that evidence be elicited from examinees?" Tasks are designed with the Student Model in mind. What features are needed to elicit the type of evidence needed?
Assembly Model	Specifications or strategies for choosing combinations of tasks to administer to an examinee.	How the test is assembled bears on its statistical value. What KSAs are elicited through a particular set of tasks? Both the effective definition of Student Model Variables and the precision with which they will be measured are shaped by the assembly model.

A *template* is a second-layer abstraction that contains more specific and technical information about the interrelations of Examinee, Evidence, and Task Models, in terms of the details of psychometric models, stimulus materials, evaluation algorithms, and so on that instantiate an assessment argument.

The CAF imposes what database designers call "normalization." It provides a unique field within a unique record type in which to place every bit of information that will be needed to construct an assessment, and discourages the duplication of information. Using the *template* form requires defining an Activity (typically based on a task or group of tasks) that is composed of a specific Measurement Model, Evaluation Procedures, Work Product(s), Materials and Presentation, Presentation Logic, and Activity-level Task Model Variables. If the Measurement Model is one of the MRCML families of models, the Measurement Model includes a definition of model type (e.g., dichotomous, partial credit), an Observable Variable, Student Model Variable(s), a Scoring Matrix, and a Design Matrix. The Evaluation Procedures include at least one Evaluation Phase in which Work Products, Task Model Variables, Input Observable Variable(s), Output Observable Variable(s), and Evaluation Action Data (e.g., information needed to evaluate Output Observable Variables from examinee Work Products[7]) are specified. When a *template* is sufficiently complete to specify a particular task, it is called a *task specification*, a blueprint that will serve the user of an authoring system in the creation of an actual assessment task. *Task specifications* reference only a single Student Model, and the Task Model Variables and Materials and Presentation choices must be fixed within a *task specification*. All of the assessment components defined in the CAF layer provide the foundation for *assessment implementation*.

The PADI team developed a stand-alone scoring engine to work in conjunction with the PADI design system. Following an assessment delivery event, the Scoring Engine sorts out evidence and provides scores based on examinee proficiencies. This Scoring Engine, called the Bear Scoring Engine, is invoked via the Internet using an XML protocol. Based on the work of Wilson and his colleagues (e.g., Adams, Wilson, & Wang, 1997) with multivariate psychometric models, the BEAR Scoring Engine uses multidimensional random coefficients multinomial logit models (MRCMLM) to accommodate more complex measurement tasks involving multidimensionality, partial credit, rating scales, and conditional independence. The Mystery Powders exemplar employs a two-dimensional MRCML model, described in a later section, with bundling to handle conditional dependence among Observable Variables. The Scoring Engine, like the other data structures within PADI, is presented as an extensible object model that can accommodate a family of psychometric models and meet varied assessment purposes. The PADI object model is designed to be extensible to psychometric models lying outside the MRCMLM family, such as the three-parameter logistic IRT model, latent class and factor analysis models, and Bayes nets (Almond & Mislevy, 1999).

In the *assessment implementation* layer, the focus is on the creating the actual objects and processes that will be used in the assessment, according to the specifications that

[7] Observable Variables are outcomes of Evaluation Phases. This is typically evaluated features of Work Products, such as a list of item scores or grades on a task. More complex Evaluation Phases, such as the one used in Mystery Powders, carry out additional steps of combining or collapsing information from earlier Evaluation Phases, so that Observable Variables themselves can be input. In automated scoring of essays, for example, multiple stages of evaluation are required to first identify syntactic and lexical features and then combine them to best approximate human raters (Shermis & Burstein, 2003).

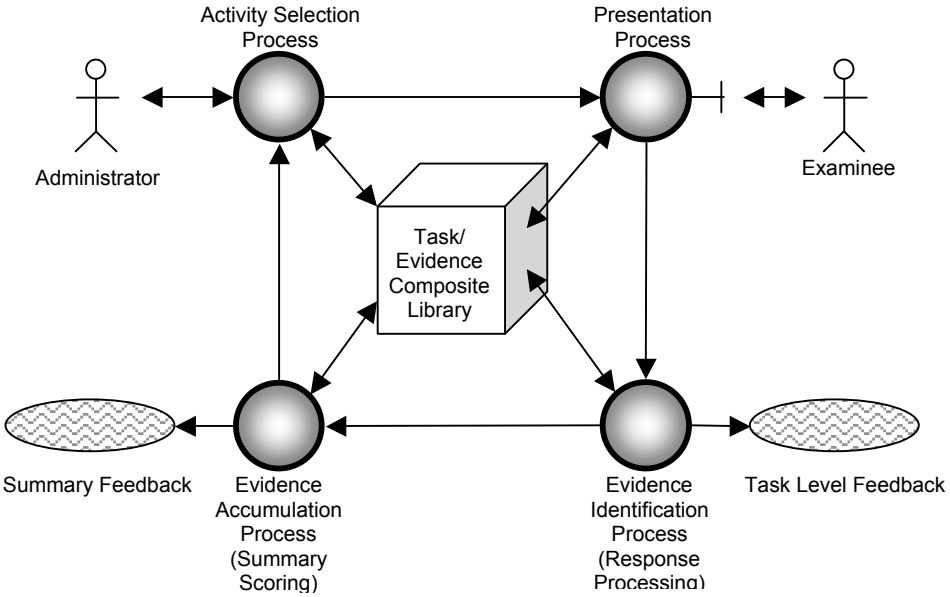
emerge from the CAF layer. These activities include task authoring[8], fitting psychometric models, estimating item parameters, assembling assessment booklets for paper-and-pencil tests, and rendering tasks for computer-based assessment delivery. Any number of tasks may be created from *templates*. While tasks may vary in their surface characteristics (for example, the specific laboratory materials for a hypothetico-deductive problem-solving task), a common CAF layer assures a common rationale and assessment argument (Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002).

The last layer of assessment design depicted in Figure 1 is *assessment delivery*. The primary activities that occur in this layer are represented in the four-process delivery system - the phase during which the operational processes of the assessment are put into action. Delivery phase elements of Mystery Powders application are shaded in light gray in Figure 1. Those outside of the scope of the Mystery Powders example are shaded in dark gray.

Figure 2 shows the flow of the design for the elements that support the four-process delivery architecture (Mislevy, Almond, & Steinberg, 2002; Almond, Steinberg, & Mislevy, 2002) as used by Mystery Powders. This four-process architecture for *assessment delivery* consists of (1) *activity selection*, (2) *activity presentation*, (3) *evidence identification* or response processing, and (4) *evidence accumulation* or summary scoring. A central database library serves to tie together information from each process. Examinees interact solely with the *presentation process*; the other processes are typically invisible to the examinee. Figure 2 presents the architecture pictorially. The *activity selection process* selects a task (or, more generally, a set of tasks or other activities) and then directs the *presentation process* to display it. When the participant has finished interacting with the task, the *presentation process* sends the results (a Work Product) to the *evidence identification process* (response processing). This process evaluates essential Observations about the results and passes them to the *evidence accumulation process* (summary scoring) which updates the Scoring Record, which updates the record of what is then known about the student. All four processes contribute information to the results database. The *activity selection process* then makes a decision about what task to present next, based on these system's current beliefs about participant ability and other criteria.

[8] It is not within the scope of the PADI project to create an authoring system, but it is within the scope of the project to provide supports for psychometric modeling.

Figure 2. The Four-Process Architecture for Assessment Delivery



Adapted from Mislevy, Almond, & Lukas, 2004

2.0 Reverse Engineering Mystery Powders

Mystery Powders is typically a hands-on lab experiment for middle school science participants. The examinees are given a white powder to evaluate, usually a mixture of two or three white substances. Examinees observe the outcomes of applying laboratory processes and reagents to the white powder and use their observations to deduce its composition.

Mystery Powders experiments are typically assessed according to some combination of examinees' lab notes, lab techniques, abilities to work collaboratively, reasoning processes, and conclusions. These examinee work products are consistent with the slow tempo of a lab-based assessment.

shows a typical worksheet and a portion of the rubric used to score it, in this case from the Stanford SEAL project (Stanford Education Assessment Laboratory, 2004). In this version of the assessment, the examinee is asked to write a complete description of what he or she has observed.

The Mystery Powders example was conceived of as an illustration of the four-process delivery architecture in coordination with major PADI components (e.g. the PADI object model). Features desired for the demonstration included electronic task representation, response, and scoring. Our first job towards this end was to create an assessment blueprint from which to create a computer-based version of the Mystery Powders task. We accomplished this through reverse engineering of extant Mystery Powders assessments, which we will describe presently. Through reverse engineering, we developed our understanding of the various layers of assessment design.

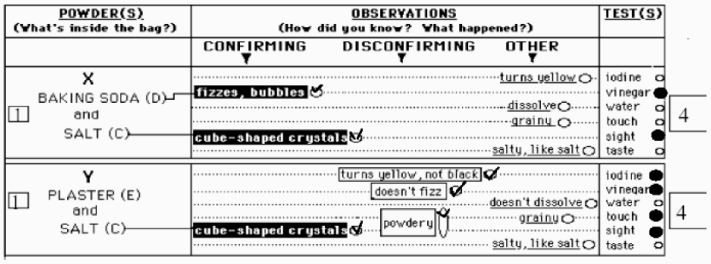
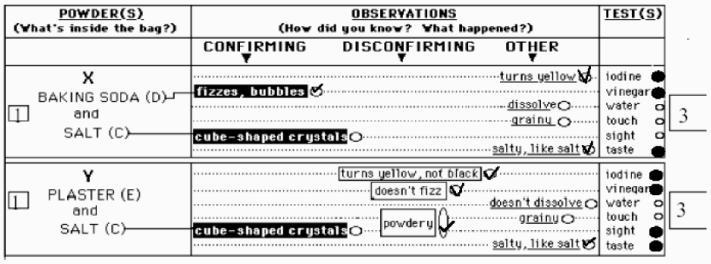
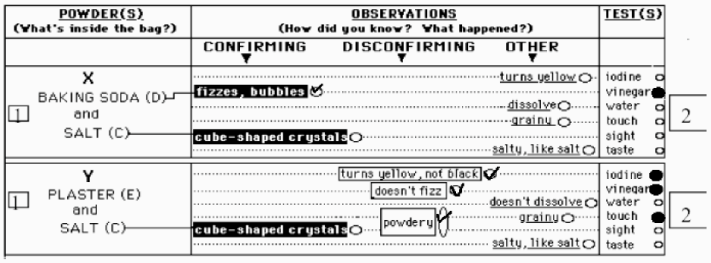
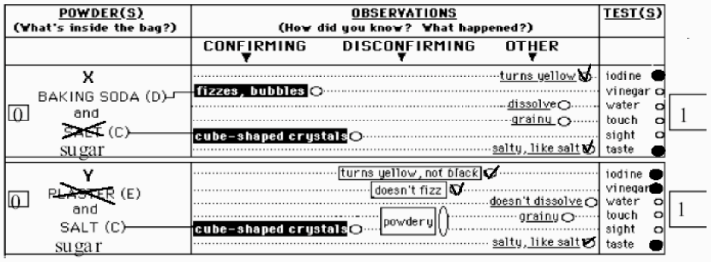
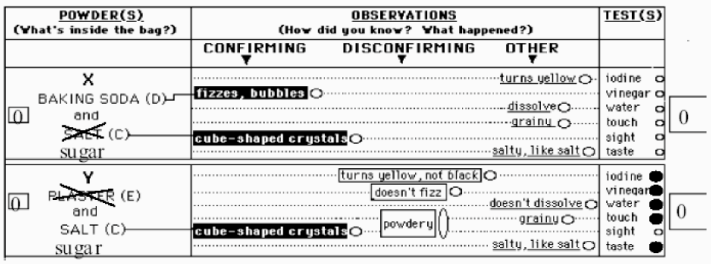
Reverse engineering is the process of discovering and articulating the technological principles of an application through analysis of its structure, function, and operation. IEEE (2003) defines reverse engineering as the process of creating a design or blueprint by analyzing a final product or system—often via identification of system components and their interrelationships—and creating representations of that product or system in an enhanced form or at a higher level of abstraction.

In this project, evidence–centered design (ECD) is the lens through which the reverse engineering and analysis work is carried out. In reverse engineering an already-existing task using the PADI design system, the given task is parsed according to the attributes of the assessment objects that compose the Examinee, Evidence, and Task Models. Such parsing requires in-depth analysis of the task and results in a "trace" of the analysis work—a PADI representation in the form of a *design pattern*, *template*, or *task specification*.

In reverse engineering, we analyzed a number of Mystery Powders laboratory tasks:

- Mystery Powders, University of Montana (http://www.eduref.org/Virtual/Lessons/Science/Chemistry/CHM0200.html)
- Mystery Powders, 5th Grade, Utah Education Network (http://www.uen.org/Lessonplan/preview?LPid=2176)

Figure 3. Stanford SEAL (2004) Mystery Powders Student Worksheet and Scoring Rubric Examples

Quality of Evidence	Example
Complete Evidence: 4 Points All **black** and all **white** observations and corresponding tests are reported.	
Strong Evidence: 3 Points One **black** *and* one or more **underlined** *and/or* one or two **white** observations and corresponding tests are reported.	
Partial Evidence: 2 Points One **black** *or* three **white** observations and corresponding tests are reported.	
Weak Evidence: 1 Point One or two **white** *and/or* one more **underlined** observations and corresponding tests are reported.	
Inadequate Evidence: 0 Points No relevant tests and observations reported *or* tests are reported without any observations.	

- Mystery Powders, Gail Baxter, CRESST Report #398, (http://www.cse.ucla.edu/CRESST/Reports/TECH398.PDF)
- Magic Powders, 5th Grade, Utah Elementary Core Curriculum Standards (http://www.westminstercollege.edu/education_gslp/brenda_julianne.pdf)
- The Scientific Method using Mystery Powders, Chicago High School for Agricultural Sciences (http://www.iit.edu/~smile/ch9305.html)

Preliminary analyses revealed a number of commonalities among the tasks described in these instantiations:

- The classic Mystery Powders task is designed to require one class period in a laboratory setting. Several of the reactions, such as the hardening of plaster of Paris and the color change of iodine from brown to blue (in the presence of starch), take a fair amount of time.
- Most of the assessments provide instructions to the teacher on how to prepare the assessment, including lists of materials, ahead-of-time setup procedures, examples of forms for the examinees to fill out, and of scoring rubrics.
- The examinee work products are consistent with the slow tempo of a lab-based assessment. Typically, examinees use worksheets to write descriptions of their observations (see, right hand column).
- Examinees' work is typically evaluated for completeness and accuracy via a rubric (again, see Figure 3)

From our analyses of the available Mystery Powders experiments, we made a number of observations, summarized in Table 2.

Table 2. Reverse Engineering Analysis of Different Mystery Powders Experiments

1. The setting of the experiment	The Mystery Powders experiment is always set in a laboratory. None of those we analyzed were computer-delivered, pencil-and-paper, or thought exercises.
2. The number of powders available	Between two and five powers were potentially included in the mixture. We deduced that the following were constraints on the selection of powders: ▪ The powders and reagents have to interact within a short time to fit the constraints of a half-hour lab. ▪ Nothing poisonous can be used. ▪ If taste is one of the tests, the powders have to be common foodstuffs. ▪ Because the focus is more on laboratory technique than deductive logic, the experiments are designed so that examinees who execute their tests correctly will probably deduce what is in the mixture.
3. The number of powders in the experiment	The median number of powders actually used in a mixture is two.
4. The time allotted to the experiment	It appears that examinees typically are given one mixture of powders to solve in one lab period. Stanford SEAL (2004) gives 25 minutes for the experiment.
5. The expected examinee work products	The typical work product is lab notes. Some versions of the task use pre-printed observational forms. The teacher typically is expected to make observations about the examinees' laboratory technique and engagement in collaboration.
6. The scoring rubrics for the work products	The rubrics typically target accurate observations and deductions about the makeup of the Mystery Powder. The rubrics also target the completeness of the description, working well in groups, adhering to safety instructions, and general lab technique.

The terminology of evidence-centered design, on which PADI is based, provides a systematic means of analyzing several of the classic Mystery Powders experiments. We concluded that the Knowledge, Skills, and Abilities (KSAs) being assessed were, in rough order of significance: (1) domain knowledge (correct interpretation of test results), (2) thoroughness of laboratory notes, (3) laboratory procedures, and (4) group work. We found that the Work Products (examinee-produced evidence) included lab notes and

write-ups. The Observed Variables were assessments of examinees' Work Products, and ratings of lab technique, collaboration and other behaviors. Different experiments used different rubrics to evaluate their Observed Variables.

In none of the existing assessments that we examined were all of the above variables explicitly defined. In particular, we had to deduce the Knowledge, Skills, and Abilities the experiment sought to measure – the precursors of Student Model Variables in a fully detailed assessment. We could not find any discussion of the relative difficulty of solving different Mystery Powder combinations or of how to assign Mystery Powders to examinees. We did not encounter any language describing Measurement Models or attempts to identify an examinee's true score on any dimensions of underlying construct(s). There was no consideration of adaptive testing because only one task was anticipated in each of the assessments.

3.0 Domain Analysis

The following sections walk through the development of MP-QTI, one layer at a time. The process by which we constructed the *domain analysis* layer reflects the constraints imposed by our target assessment vehicle and by "reverse engineering." It had been determined a priori that a computer-based, automatically-scored assessment was needed to demonstrate the relationships between PADI design objects and the four-process delivery architecture, and that an adaptation of the well-known hands-on performance assessment task Mystery Powders might be suitable. Reverse engineering required understanding how the Mystery Powders assessment is typically administered, to better understand the science, the assessment approach, and the design decisions that led to its current form. Viewing this information in a broader context (specifically, hypothetico-deductive reasoning in a finite solution space), we determined how to build on the key reasoning aspects and basic activities of Mystery Powders, but make design decisions that would fit the constraints of the intended demonstration. This forward engineering activity is subsequently described in the *domain modeling* layer, the creation of a PADI *template* (CAF layer), and the assessment instantiation in an operational delivery system.

Domain analysis, in the context of implementing a demonstration of four-process delivery of the well-known Mystery Powders experiment, was more abbreviated than it would have been for an assessment that was either new or intended to be delivered to real examinees. Our approach was different from that which a real assessment would have demanded. A wholly new test might have involved analyzing several domains with the intent of constructing an assessment that involved: (1) inferential reasoning from the domain of cognitive science, (2) the domain that chemists and physicists might call properties of matter - chemical reactions and laboratory technique, and (3) the curricular and instructional domain within which the students for whom the assessment was intended study. Instead, we worked backwards, reverse engineering existing instances of Mystery Powders exercises in the attempt to infer what knowledge and skills it was that they were developing or assessing. Most of the instances we found were characterized as laboratory experiments—teaching exercises rather than assessments. Only a few included rubrics, suggesting that they might not have been designed to be graded.

Given this context of assessment development, we reached some understandings of the domains Mystery Powders is drawn from and implications of the understandings for assessment. Mystery Powders is an example of a class of hypothetico-deductive reasoning problems that appear in many domains (e.g., troubleshooting mechanical or electronic systems). Specifically, Mystery Powders involves problem-solving within a finite solution space, in this case using content knowledge from the domain called properties of matter. For tasks involving problem-solving within a finite solution space (Table), examinees are presented with a problem of determining the state of an object or system and methods for gathering information about its state. No method is definitive; the observations yielded by any method can be used to rule in some possibilities and rule out others. Effective problem solution requires examinees to think and reason with subject matter knowledge. The nature and quality of cognitive activity underlying an individual's performance (e.g., problem representation, solution strategies, solution monitoring,

explanations) reflects the experience, degree of learning, and state of knowledge of the problem solver. In particular, effective hypothetico-deductive problem solving requires an understanding of the procedures that can be applied to rule in or out particular states, being able to interpret the results of any tests that are applied, synthesizing information to determine what states are still possible at different points in the process, and being able to choose new tests that will effectively narrow the search space.

Because Mystery Powders also is drawn from the domain referred to as properties of matter, what is required to solve hypothetico-deductive tasks in this instance will necessarily involve content knowledge of chemistry (e.g., chemical reactions) as well as knowledge of laboratory techniques. The level of content knowledge required for hypothetico-deductive tasks generally can range from little or no content knowledge (e.g., “20 Questions” and Milton Bradley’s “Guess Who” game) to complex and specialized content knowledge (e.g., finding the fault in the ill-fated Challenger space shuttle (House of Representatives, Committee on Science and Technology,1986).

In educational contexts, we can consider task use along a number of dimensions. Tasks may be used for learning purposes, devoid of any formal assessment or grading procedure. Tasks may be used for formative assessment purposes in instruction or self-assessment – to improve a teacher’s instruction or to assist examinees in gaining deeper understandings of their own knowledge, skills, and abilities. Tasks may also be used for summative assessment purposes – administered under standardized conditions, perhaps at a large-scale, and providing reliable, valid evidence of what examinees know and can do. Evidence of problem-solving can include written or verbal descriptions, observations of examinee actions, and some record or solution spaces as the examinee(s) proceeds through the task. Tasks’ contexts may be authentic or virtual - involving laboratory materials, simulated demonstrations, talk-alouds, or some combination thereof. Tasks may be presented for individual or group problem-solving. Such considerations move us toward defining the general components of an assessment argument through domain modeling.

Table 3. A General "Hypothetico-Deductive Problem Solving in a Finite Space" Design Pattern Related to Mystery Powders

Title	Hypothetico-Deductive Problem-Solving in a Finite Space	
Summary	Examinees are presented with a problem of determining the state of an object or system, and methods for gathering information about its state. No method is definitive; each rules in some possibilities and rules out others. Effective problem solution requires examinees to think and reason with subject matter knowledge. The nature and quality of cognitive activity underlying an individual's performance (problem representation, solution strategies, solution monitoring, explanations) reflects the experience, degree of learning, and state of knowledge of the problem solver.	The system and tests may be real and carried out hands-on, or virtual as in a simulation or talk-aloud solution. The emphasis in this *design pattern* is on procedures and strategies.
Focal Knowledge, Skills, and Abilities	Ability to apply content knowledge to solve a problem. Ability to generate and elaborate explanations of task-relevant concepts. Ability to build a mental model or representation of a problem to guide solution. Ability to manage thinking during problem-solving and to allocate resources efficiently. Ability to enlist appropriate goal-directed solution strategies.	
Rationale	Examinees' ability and inclination to solve problems effectively depend on their having certain knowledge, skills, and attitudes. Knowledge has limited problem-solving value in the absence of knowing when and how to apply that knowledge (i.e., integrated knowledge). Integrated knowledge structures, characteristic of effective problem solvers, are displayed in the ability to represent a problem, select and execute goal directed strategies, monitor and adjust performance, and offer complete, coherent explanations. In particular, problem-solving to determine the state of a finite system with a set of tests requires an understanding of the procedures that can be applied to rule sets of states as in or out, being able to interpret the results of the tests, synthesizing their information to determine what states are still possible after a series of tests, and being able to choose a next test that will effectively narrow the search space.	
Additional Knowledge, Skills, and Abilities	Domain knowledge	Hypothetico-deductive problems in finite spaces may address systems that require very little domain knowledge (e.g., Milton Bradley's "Guess Who?" game for children) or be very complex and require specialized knowledge (e.g., finding the fault in the hydraulics system of an F-15 aircraft).
	Capability to carry out tests	When the task is physical, carrying out the test procedures may itself be require knowledge and proficiency (e.g., lab tests in chemistry)
	Ability to work with a group	Tasks may be carried out by a group rather than an individual, in which case *design pattern*s useful to assessing group work will also be consulted for task design.
Potential observations	Correctness of answer to problem and quality of evidence to support that answer/conclusion Quality of explanation of task-specific concepts Adequacy of problem representation or problem-solving plan Appropriateness of solution strategies Frequency and flexibility of self-monitoring Quality of evidence to support a particular answer/conclusion Power of selected tests Accuracy of deductions at each step of each solution	

Potential work products	Written or verbal description/identification of where the problem is or what the solution is to the problem.	Answer.
	Illustration of problem solution and/or written justification for "Here's how I know."	Evidence.
	Verbal or written description of anticipated problem-solving approach.	Problem representation or plan.
	Verbal or written explanation of task-specific concepts.	Explanation.
	Log or observation of examinee actions.	Record of actions made automatically or by an observer--i.e., not the problem-solver. In particular, contains the sequence of test procedures that were carried out.
	Observation data / log file (to some extent) / think aloud protocols of what examinee attends to or thinks about while solving the problem.	Monitoring.
	Indication of which possibilities are ruled in or out by a given test procedure.	
	Indication of which possibilities are ruled in or out by all test procedures given thus far, at any given point during the solution.	
Potential rubrics	Answer key	Generally dichotomous (right, wrong)
	Evidence key	Degree of completeness and relevance of evidence ranging from irrelevant or inadequate to complete.
	Explanation Rubric	Qualitative levels of task-specific conceptual knowledge as expressed in explanations ranging from single statement of fact to accurate, coherent, and complete.
	Problem Representation Rubric	Qualitative levels of proposed solution plans ranging from virtually non-existent ("I'll just try this") to articulation of a reasoned, coherent, set of actions and anticipated outcomes.
	Strategy Rubric	Qualitative levels of implemented solution strategies ranging from undirected trial and error to efficient, informative, goal oriented. May include consideration of sequence.
	Monitoring Rubric	Qualitative evaluation of frequency and flexibility of self-monitoring of content knowledge, task constraints, and interpretations of current findings.
	Match of current hypothesis to what is potentially known.	After a given sequence of tests, it is possible to calculate what possibilities are in and out. This vector is compared with the examinee's belief at that point, with higher values indicating better match.
Characteristic features	Statement of problem provides system, initial conditions, and set of test procedures.	
	System with imperfectly known state (e.g., fault, unknown components)	Virtual or physical
	There is a finite (though possibly large) space of possibilities of the system state.	Thus the problem can be definitely solved. If an open-ended problem space is desired instead, see the *design pattern* for hypothetico-deductive problem-solving in an open system.
	Each test procedure rules some aspects of system state in and others out.	No single test is definitive, so that the examinee must carry out multiple tests to arrive at a solution.

Variable features	Level and nature of content knowledge required to solve problem.	Level -> depth Nature -> subject matter See Baxter/Glaser content-process space
	Degree of scaffolding or prompting.	
	Individual work, with a partner, or as a member of a group?	
	Degree of domain familiarity.	If may be that the assessor knows the domain is familiar to the examinee, knows it is unfamiliar, or does not know. If it is known to be familiar, examinee's problem-solving reflect mainly strategy & procedures; if the assessor does not know the examinee's familiarity, then domain knowledge as well as problem-solving in the domain are sources of variation in examinee's performance.
	Number of variable features in unknown system.	More features & possibilities make the problem more difficult.
	Number of tests to choose from.	Larger set of choices makes the problem more difficult.
	Redundant tests?	Presence of redundant tests makes the problem more difficult.
	Overlapping tests?	Problem is easier if tests are orthogonal in the information they provide.
I am a kind of	Conduct Investigations Implement Solution Strategies Modifying solution strategies based on external feedback, self-monitoring, and reflection Problem Solving	
These are kinds of me	Mystery Powders	
These are parts of me	Generate explanations based on underlying scientific principles. Implement solution strategies. Modifying solution strategies based on external feedback, self-monitoring, and reflection. Monitoring strategies. Plan systematic solution strategies. Use data to support scientific argument.	
Educational standards	Unifying Concepts 1.2. Evidence, models, and explanation NSES 8ASI1.2. Design and conduct a scientific investigation. NSES 8ASI1.3. Use appropriate tools and techniques to gather, analyze, and interpret data.	
Exemplar tasks	Mystery Powders IMMEX problem solving tasks: http://www.immex.ucla.edu/	
Online resources	http://www.stanford.edu/dept/SUSE/SEAL/assessments/Assessments.htm http://www.cse.ucla.edu/CRESST/Reports/TECH398.PDF	
References	Baxter, G.P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. Educational Psychologist, 31(2), 133-140.	Cognition and Assessment
	Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 17(3), 37-45.	Cognition and Assessment
	Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). The nature of expertise. Hillsdale, NJ: Erlbaum.	Expert-Novice Research on Problem Solving
	Ericsson, K. A., & Simon, H. A. (Eds.). (1991). Toward a general theory of expertise: Prospects and limits. New York: Cambridge Press.	Expert-Novice Research on Problem Solving
	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.	Description of the HYDRIVE intelligent tutoring system, which is built around hypothetico-deductive reasoning for troubleshooting the hydraulics system of the F-15 aircraft.

4.0 Domain Modeling

4.1 Domain Modeling Overview

Domain modeling is the process of organizing the considerations identified in *domain analysis* into an assessment argument. Mystery Powders tasks are a particular instance of a broader class of science inquiry assessments called hypothetico-deductive problem-solving tasks in a finite space. In this description of the *domain modeling* layer for Mystery Powders, we will discuss a broader, motivating *design pattern* for hypothetico-deductive tasks (Table 3), then and consider a more specific *design pattern* for Mystery Powders tasks (Figure 5). Finally, we will discuss a rapid prototype in Excel for MP-QTI to check assumptions, obtain additional information, and sharpen our design choices before starting in to the more formal design elements of the *conceptual assessment framework*.

4.2 The Motivating Design Pattern

As mentioned earlier, for tasks involving hypothetico-deductive problem-solving within a finite solution space (see Table 3), examinees are presented with a problem of determining the state of an object or system and methods for gathering information about its state. The Focal Knowledge, Skills, and Abilities required for tasks within this class include the abilities to apply content knowledge to solve a problem, generate and elaborate explanations of task-relevant concepts, build a mental model or presentation of a problem to guide solution, manage thinking during problem-solving and allocate resources efficiently, and enlist appropriate goal-directed solution strategies. Key features of this class of tasks include: (1) a statement of problem that provides a system, initial conditions, and set of test procedures, (2) a system with an imperfectly known state, (3) a finite (though possibly large) space of possibilities within the system state, and (4) test procedures that increasingly restrict the solution space.

There are any number of types of tasks that belong to this class of hypothetico-deductive problem-solving within a finite space, Mystery Powders tasks being but one example. Tasks as diverse as the Milton Bradley children's game *Guess Who?* and the aircraft hydraulics troubleshooting problems in the Hydrive coached practice system (Steinberg & Gitomer, 1996) also belong to this class. This illustrates the intent of *design pattern*s that are based on paradigmatic approaches to gathering evidence for assessing key aspects of scientific reasoning. These approaches are captured in the Characteristic Features attribute of *design patterns*. These tasks can vary considerably as to content demands, difficulty, type of response, and so on. Organized around the structure of an assessment argument, a *design pattern* is meant to help designers think about these issues from early on in the design process.

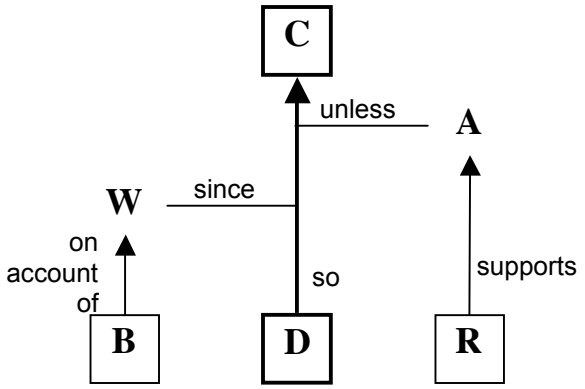
4.3 The Assessment Argument

Evidence-centered design uses the term "assessment argument" to define the chain of reasoning that links the evidence that an examinee provides in response to an assessment context and goal with inferences about the examinee's Knowledge, Skills, and Abilities (KSAs). PADI has adapted structures called *design patterns* to help organize information from *domain analysis* into the form of potential assessment argument (PADI,

2003). An assessment *design pattern* helps domain experts and assessment designers fill in the slots of an assessment argument. Because the structure of a *design pattern* implicitly contains the structure of an assessment argument, filling in the slots simultaneously renders explicit the relationships among the pieces of information in terms of the roles they will play in the argument. We can speak of the assessment structure provided by the *design pattern* and the assessment substance as determined by the assessment designer (Mislevy, 2003).

The concern of evidence-centered assessment design in the layer of *domain modeling* is to lay out an assessment argument schema. Toulmin's general structure for arguments, in terms of claims, data, and warrants, provides a starting point (Figure 4). In an assessment argument, the claim refers to the target of the assessment, such as the level of proficiency in scientific problem-solving or the ability to use language correctly in varying contexts. There are two kinds of data that support claims about students: the features of the assessment task, which are data the task designer is mainly responsible for providing, and the features of the student's work in such settings. The warrant is the logic of the reasoning that explains why certain data should be considered appropriate evidence for certain claims. The work carried out in a *domain analysis* (in our case, reverse engineering a variety of existing Mystery Powders tasks) provides the raw material for assembling these elements.

Figure 4. Toulmin's (1958) Structure for Arguments

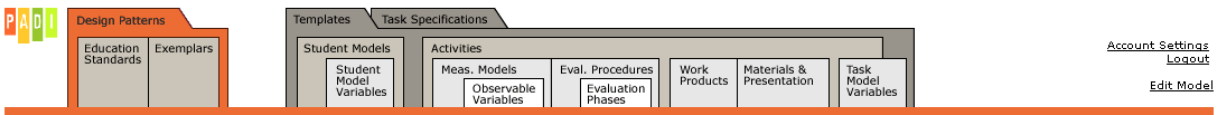


Reasoning flows from *data* (D) to *claim* (C) by justification of a *warrant* (W), which in turn is supported by *backing* (B). The inference may need to be qualified by *alternative explanations* (A), which may have *rebuttal evidence* (R) to support them.

4.4 A Design Pattern for Mystery Powders

The second, more specific *design pattern* that was created for Mystery Powders tasks (see Figure 5) can now be considered in terms of the assessment argument. This *design pattern* limits hypothetic-deductive problem solving tasks (in a finite space) to those involving unknown combinations of known white powders and a set of physical and chemical tests. Examinees are told the entire set of potential powders in a given mix and instructed to select a sequence of tests from a finite and pre-specified set of tests. They potentially may be scored on a variety of variables that could include the correctness of their solutions, their test selection strategy, the accuracy of their deductions following each test, their choices of evidence, their lab techniques, and their group participation.

Figure 5. A *Design Pattern* for Mystery Powders



Mystery Powders - Generalized | Design Pattern 1975

[View Tree | Export]

Title:	Mystery Powders - Generalized	
Summary	This design pattern is for potential Mystery Powders assessments targeted to middle- or high-school students. The student is presented with a mystery powder, or series of powders, the composition(s) of which the student must determine, given a finite number of potential powders in a mix. The student selects from chemical and physical tests that can be performed on a mixture, and either observes the test results or is told what observations are available from the test results. No method is definitive; each rules in some possibilities and rules out others. An effective problem solution requires the student to think and reason with subject matter knowledge. The nature and quality of cognitive activity underlying an individual's performance includes problem representation, solution strategies, solution monitoring, and explanations. The system and tests may be real and carried out hands-on in a laboratory setting, or virtual as in a simulation or talk-aloud solution.	
Focal Knowledge, Skills, and Abilities	Domain knowledge	Of chemistry relating to potential powder components and their reactions with various reagents
	Ability to apply content knowledge to determine the composition of a mystery powder.	
	Ability to generate and elaborate explanations of why a particular test was chosen and how the presence/absence of a particular powder was determined.	
	Ability to build a mental model or representation of a problem space to guide solution.	Here particularized to the content and context of the universe of mystery powders tasks. Building and using mental models always involves, and is therefore in a sense conditional on, some particular model and context.
	Ability to manage thinking during problem-solving and to allocate resources efficiently.	e.g., hypothetical reasoning - what will happen in a given test if a certain powder is in or out of a mixture. Here particularized to the content and context of the universe of mystery powders tasks.
	Ability to enlist appropriate goal-directed solution strategies, such as strategic efficiency in choosing tests.	Here particularized to the content and context of the universe of mystery powders tasks.
Rationale	Students' ability and inclination to solve scientific problems effectively depend on their having certain knowledge, skills, and attitudes. Knowledge has limited problem-solving value in the absence of knowing when and how to apply that knowledge (i.e., integrated knowledge). Integrated knowledge structures, characteristic of effective problem solvers, are displayed in the ability to represent a problem, select and execute goal directed strategies, monitor and adjust performance, and offer complete, coherent explanations. In particular, problem-solving to determine the state of a finite system, such as the composition of a mystery powder, with a set of tests requires an understanding of the procedures that can be applied to rule sets of states as in or out, being able to interpret the results of the tests, synthesizing their information to determine what states are still possible after a series of tests, and being able to choose a next test that will effectively narrow the search space.	
Additional Knowledge, Skills, and Abilities	Capability to carry out tests	When the context for the mystery powders task is lab-based, carrying out the tests may itself require additional knowledge of lab-based tests (e.g., how long to wait for reactions) and physical skills (e.g., dexterity).
	Ability to work within a group	Tasks may be carried out by a group rather than an individual, in which case design patterns useful to assessing group work will also be consulted for task design.
	Deductive reasoning	based on assuming certain powders are included or excluded from a mixture, determining which other powders must be in or out, given the minimum and maximum number of powders in the mixture (part of the problem situation description)
Potential observations	Correctness of identification of a given powder	
	Power of selected tests (how much new information each uncovered)	Some tests provide more information than others at a given point in solving a given mixture.
	Accuracy of deductions at each step of a solution	e.g., following each test, how accurate were specifications of which powders were in / out / or unknown?
	Presence of particular observations for a given test/powder	e.g., when salt is present and a taste test is given, observations include "salty taste"
	Quality of participation in collaborative work.	Relevant when tasks are solved by groups. See "Participate in collaborative scientific inquiry" design pattern.

Figure 5. A *Design Pattern* for Mystery Powders (continued)

	Appropriateness and quality of laboratory techniques	e.g., amount of time given to particular test. Relevant with hands-on solution, as opposed to computer simulations and hypothetical talk-alound solutions.
	Quality of students' representation of the entire problem to be solved	
	Quality of explanations for why a particular test was chosen and how the presence/absence of a particular powder was determined	
Potential work products	Representation of problem to be solved	Written and/or verbalized; language-based and/or diagramatic
	Selection of (and sequences of) tests	Sequences of tests may not be provided on some versions of the assessment
	Observations made during a task	Verbal observations, lab notes, etc.
	Deductions made about presence/absence of powders after each test	
	Final solution for a given mixture	
	Discourse and behavior with other group members during task	Relevant when tasks are solved by groups. See "Participate in collaborative scientific inquiry" design pattern.
	Explanations for why a particular test was chosen and how the presence/absence of a particular powder was determined	Written and/or verbal
	Lab-based actions using experimental techniques	Individual, partnered, group-based
Potential rubrics	Answer key for final solution to given mixture	Often dichotomous (right, wrong), but can be partial credit for number of powders identified correctly as in, out, or indeterminate.
	Strategy rubric.	Qualitative levels of implemented solution strategies ranging from undirected trial and error to efficient, informative, goal oriented. May include consideration of sequence. See Baxter et al reference for example.
	Match of current hypothesis to what is potentially known.	After a given sequence of tests, it is possible to calculate what possibilities are in and out. This vector is compared with the student's belief at that point, with higher values indicating better match.
	Evidence key	Degree of completeness and relevance of evidence ranging from irrelevant or inadequate to complete.
	Participation rubric(s)	Degree of participation ranging from disengaged to fully contributing to group task and discussion, and exhibiting role flexibility (e.g., recorder, leader, observer)
	Lab-based techniques rubric	Qualitative levels of use of laboratory techniques ranging from undirected to completely correct use of equipment (appropriate, safe, clean)
	Problem Representation Rubric	Qualitative levels of proposed solution plans ranging from virtually non-existent ("I'll just try this") to articulation of a reasoned, coherent, set of actions and anticipated outcomes.
	Explanation rubric	Qualitative levels of task-specific conceptual knowledge as expressed in explanations ranging from single statement of fact to accurate, coherent, and complete.
Characteristic features	Statement of problem provides potential powders for any given mixture, initial conditions, and set of test procedures	
	System with imperfectly known state (e.g., composition of given powder)	
	There is a finite (though possibly large) space of possibilities of the system state	
	Each test procedure rules some aspects of system state in and others out	
Variable features	Setting / administration mode for mystery powders task	Lab-based, simulation, think-aloud
	The number and types of powders in a given mixture	
	The number and types of potential powders in a mixture	
	The number and types of potential tests (e.g., chemical and physical) for a mystery powders task	
	The presence of and settings for the minimum and maximum numbers of powders in a given mixture (part of problem description situation)	These constrain the solution space
	Inclusion of indeterminant powders	Whether mixtures are given in which the available tests do not produce enough information to conclusively determine the solution
	Inclusion of ambiguous tests	Test results themselves are not conclusive (e.g., color of iodine test is not clearly blue)
	Degree to which a test procedure efficiently splits the solution space (rules out some possibilities and rules in others)	
	Individual, partnered, or group-based task	e.g., group-administered lab-based task
	Level of scaffolding for the task	e.g., specificity of lab-based instructions, guidance for interpretation of test results

Figure 5. A *Design Pattern* for Mystery Powders (continued)

		Level of demand of content knowledge	Depth and breadth of content knowledge need to solve a particular task
		Visual presentation of test results	For a simulated version of Mystery Powders assesment
		Use of repeated tasks - i.e., different mystery powders to solve	e.g., for simulated version of the test. For learning purposes, can sequence tasks from easier to harder and/or more to less scaffolding.
		Use of adaptive selection of subsequent tasks	Possible for assessments with multiple tasks.
		Presence and length of time limit	e.g., one class period for a lab-based version of the assessment given in a classroom setting
		Presence of previous results	For assessments with multiple tasks
		Presence of active facilitator	e.g., for talk-aloud version of the task
I am a kind of		Conduct investigations. Students are presented with a scientific problem to solve or investigate and a solution strategy. Do...	
		Hypothetico-Deductive Problem-Solving in a Finite Space. Students are presented with a problem of determining the state of an object or system, and methods f...	
		Implement solution strategies. Students are presented with a scientific problem to solve or investigate. Do they use solution strat...	
		Modifying solution strategies based on external feedback, self-monitoring, and reflection. In this design pattern, students engage in self-monitoring, reflection, and apply external feedback ...	
These are kinds of me		Mystery Powders Design Pattern. This design pattern is for the Mystery Powders Assessment. The computerized assessment presents a m...	
These are parts of me		Generate explanations based on underlying scientific principles. Students are asked questions about scientific phenomena that require them to explain what they know....	
		Implement solution strategies. Students are presented with a scientific problem to solve or investigate. Do they use solution strat...	
		Modifying solution strategies based on external feedback, self-monitoring, and reflection. In this design pattern, students engage in self-monitoring, reflection, and apply external feedback ...	
		Monitoring strategies. Students monitor their actions and flexibly adjust their approach based on performance feedback.	
		Plan systematic solution strategies. Students are presented with an open-ended problem to investigate and must generate a plan for solvin...	
		Use data to support scientific argument. A student must use data, either collected or provided, to support a scientific argument. Does the s...	
Educational standards		NSES 8ASI1.2. Design and conduct a scientific investigation. Students should develop general abilities, such as sy...	
		NSES 8ASI1.3. Use appropriate tools and techniques to gather, analyze, and interpret data. The use of tools and te...	
		NSES 8ASI1.4. Develop descriptions, explanations, predictions, and models using evidence. Students should base the...	
		NSES 8ASI1.5. Think critically and logically to make the relationships between evidence and explanations. Thinking...	
		NSES 8ASI1.7. Communicate scientific procedures and explanations. With practice, students should become competent ...	
		Unifying Concepts 1.2. Evidence, models, and explanation	
Templates		Mystery Powders Template. The task is to determine the composition of a "Mystery Powder". Such powders may consist of any mix...	
Exemplar tasks		Mystery Powders. In this performance assessment students are asked to investigate which of 3 white substances (salt, ...	
Online resources		http://www.stanford.ed...	
		http://www.cse.ucla.ed...	
References		Baxter, G.P., Elder, A. D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. Educational Psychologist, 31(2), 133-140.	Cognition and Assessment
		Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 17(3), 37-45.	Cognition and Assessment
		Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). The nature of expertise. Hillsdale, NJ: Erlbaum.	Expert-Novice Research on Problem Solving
		Ericsson, K. A., & Simon, H. A. (Eds.). (1991). Toward a general theory of expertise: Prospects and limits. New York: Cambridge Press.	Expert-Novice Research on Problem Solving
		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.	Description of the HYDRIVE intelligent tutoring system, which is built around hypothetico-deductive reasoning for troubleshooting the hydraulics system of the F-15 aircraft.
I am a part of			

The attributes of the *design pattern* are connected with elements of the assessment argument. The claim of the argument (see Figure 4) corresponds to the Knowledge, Skills, and Abilities we wish to assess. The main Knowledge, Skills, and Abilities are domain knowledge, applications of content knowledge, use and explanations of test strategies, representations of the problem space, and management of thinking during

problem-solving. Additional Knowledge, Skills, and Abilities that might or might not be required in tasks, depending on how the designer chose to structure them, include carrying out the tests, working within a group, and deductive reasoning.

Data related to the claim are understood by considering relevant observations and are collected by means of evaluating examinee Work Products. For Mystery Powders, Potential Work Products include an examinee's representation of the problem to be solved, the selection of tests performed, the sequence of tests performed, observations made during a task, deductions indicated after each test, the conclusion with regard to powder composition, discourse and behavior with other group members during the task (for collaborative versions), explanations for why a particular test was chosen, and how the present/absence of a particular powder was determined, and lab-based actions using experimental techniques. Potential Observations include the correctness of identification of a particular powder, the power of selected tests, the accuracy of deductions at each step of the solution, the presence of particular observations about the outcome of a given test of a given powder, the quality of participation in collaborative work, laboratory techniques, adequacy and accuracy representation of the problem to be solved, and the quality of explanations of why a particular test was chosen and how the presence or absence of a particular powder was determined.

Examinee data must be elicited through the structures, directives, and activities of particular tasks. There are many possible forms for Mystery Powders tasks. What features must they all share, however, in order to elicit evidence about the Focal KSAs? Characteristic Features of a Mystery Powders task include a statement of problem that provides potential powders, initial conditions, and potential test procedures; a system with an imperfectly known state; a finite space of possibilities within the system state; and test procedures that rule out some possibilities and rule in others.

Beyond these Characteristic Features, tasks may then vary according to the administration mode for the task (e.g., simulated, lab-based, talk-aloud), the number of and particular powders in a given mixture, the number of and potential powders in a given mixture, the number and types of potential tests, the degree to which available tests split the solution space, the presence of and settings for minimum and maximum numbers of powders in a given mixture, the use of indeterminate powders (some combinations cannot be distinguished when tested in certain orders), and the use of ambiguous tests (the results of a test may be difficult to perceive). Other Variable Features include whether a task is individual, partnered, or group-administered; the level of scaffolding for the task; the level of demand of content knowledge; whether test results are visually presented (e.g., in a simulated assessment); the use of repeated tasks (e.g., in a simulated assessment); adaptive selection of subsequent tasks; the presence and length of a time limit; the presentation of prior results to earlier tasks; and the presence of an active facilitator (e.g., for a talk-aloud assessment). All of these choices hold implications for the level of effort expended in accomplishing the task, what knowledge is assumed or developed, security considerations, and learning opportunities afforded. A task designer's choices are shaped by the purpose, setting, and resources for the context in which a task is to be used. It is the role of the *design pattern* to bring these choices to the attention of the task designer and provide some guidance for thinking about how to make them.

The warrant, or rationale, justifies the link between data and claim. In our case, the warrant states that examinees' ability and inclination to solve scientific problems effectively depend on their having certain Knowledge, Skills, and Abilities. Knowledge has limited problem-solving value in the absence of knowing when and how to apply that knowledge (i.e., integrated knowledge). Integrated knowledge structures, characteristic of effective problem solvers, are displayed in the ability to represent a problem, select and execute goal directed strategies, monitor and adjust performance, and offer complete, coherent explanations. In particular, problem-solving to determine the state of a finite system, such as the composition of a Mystery Powder, with a set of tests requires an understanding of the procedures that can be applied to rule sets of states as in or out, being able to interpret the results of the tests, synthesizing their information to determine what states are still possible after a series of tests, and being able to choose a next test that will effectively narrow the search space (Newell & Simon, 1972).

4.5 The Spreadsheet Mockup

We chose to develop a prototype of a simulated Mystery Powders assessment to check our assumptions, obtain additional information, and develop a better understanding of our design choices before starting to define the more formal design elements of the *conceptual assessment framework*. We created a rapid prototype of MP-QTI in Excel. Our work with this new type of Mystery Powders assessment is in essence a software development exercise, structured along the conceptual lines of ECD. We developed a prototype in an Excel spreadsheet, working out the database logic, and in the process determined that a table-driven approach would be good for the full implementation. Having project members use the prototype to work through sample tasks provided insights about the features of tasks that affect their difficulty.

As mentioned previously in our discussion of reverse engineering, a computer-delivered assessment would be expected to differ significantly from a laboratory experiment. Because the examinee does not handle laboratory materials in the computer-delivered assessment, the tempo is greatly increased; therefore, we decided that the assessment should consist of a number of tasks, each of them a unique combination of powder mixture, minimum possible powders, and maximum number of powders—rather than just one combination as is universally the case in wet–lab administrations. The examinee would be assessed on a series of Mystery Powders. In doing this, the examinee selects from finite, predetermined lists of tests.

An examinee may improve his or her inquiry skills over the course of the assessment. We came to think of the computerized assessment as being similar to a video game. We considered that the examinee can enter the assessment without extensive prior domain knowledge and gain some domain knowledge through trial and error. The examinee can develop inquiry skills associated with selecting tests based on their relative power, a skill that would not be exercised in the laboratory setting.

Lastly, the computer can spare the examinee the effort of interpreting ambiguous evidence that arise naturally in lab work. For example, consider cornstarch with an iodine test. Unless the cornstarch is first dissolved in water, adding iodine tends to turn it a nondescript brown instead of a royal blue. In a laboratory setting it might be difficult to

determine, conclusively, the results of the test. However, in the online assessment, the outcome can be blue because the computer says it is blue. Similarly, a computerized assessment can assert that plaster has no taste although it probably does, and we would not ask an examinee to find out in a lab situation. It became clear that MP-QTI would tap significantly different skills than the laboratory exercise on which it is based. In particular, it would emphasize declarative knowledge within the domain, effective reasoning techniques in determining efficacious tests, and evaluating their results in terms of effect on the search space. It would place no value on visual discernment, laboratory technique, writing ability, or the ability to work as part of a lab team. Compared to a Mystery Powders laboratory exercise, then, MP-QTI would minimize requirements of the Additional KSAs concerning physically carrying out and interpreting lab tests.

Because the MP-QTI is a fundamentally different assessment than a laboratory assessment, we needed to make some decisions about scope in adapting to the online environment. Because our intention was for each examinee to interact with many different Mystery Powders (each considered a task), we chose to include six potential powders in any given mixture in the example—more powders than used in any of the referenced assessments. Six powders can form up to 62 combinations ($2^6 = 64$, minus the null set and the complete set).

We designed the assessment with six potential chemical and physical tests that the examinees could use in any sequence to determine the composition of the Mystery Powder. Vinegar and iodine are the chemical tests ("reagents") we chose; water, heat, taste, and visual inspection are physical tests[9]. There are 63 possible combinations of these tests ($2^6 = 64$, minus the null set), including using all of them. Each of the combinations of powder can be presented to the examinee with different specifications for the minimum and maximum numbers of powders. It is easier to figure out that a powder consists of, say, cornstarch and plaster, knowing that there are precisely two powders in the mix (15 possible solutions) than knowing that there are between one and five powders (62 possible solutions). The total 480 combinations of powder mixtures and specifications for minimums and maximums comprise the "universe" of Mystery Powders tasks. With "a task solution" defined as a choice of "in," "out," or "don't–know" for each of six powders, there are 3^6=729 combinations of final solutions. With 480 tasks and 63 combinations of tests, there are 30,240 task–test combination pairings. Considering that there are also 729 possible conclusions (correct and incorrect) that could be drawn from any task and combination of tests, we learned that our scoring algorithms for MP-QTI will be rather extensive, covering millions of task-test combination-conclusion sets!

Just over a quarter of the possible tasks involve indeterminate solutions. That is, there are many combinations of powders, taken in combination with minimum and maximum settings, in which the examinee cannot conclude definitively whether or not flour is part of the mixture (125 of the 480 possible tasks). For example, test results for the mixture of plaster, flour, sugar and cornstarch are identical to those for plaster, sugar and cornstarch. The examinee can only be sure of the exact combination if they know the

[9] Actually, water serves as both a chemical reagent and a physical process. Plaster reacts chemically; salt and sugar react physically.

combination cannot include four powders or must include four powders. It would make no sense to require this level of deductive logic in a middle school lab experiment that was to be executed one time. The examinees would have no opportunity to build their deductive skills through interaction with the task. For the computer-delivered assessment with multiple powder combinations, however, it is more plausible to include this level of challenge for examinees who have mastered the basics of the assessment because middle school examinees have a demonstrated ability to learn very complex behaviors (e.g., Gee, 2003). The MP-QTI thus includes a choice as to whether the examinee will be presented with mixtures that can lead to indeterminacies.

In working with the spreadsheet and pilot data, we gained information about an appropriate level of scaffolding for examinees, at least for the purposes of this illustration of technical aspects of the PADI design framework. Specifically, we arrived at the following decisions:

- The examinee would be shown the results of all prior experiments rather than requiring the examinee to rely on his or her memory of prior results.
- After each test, the previous settings of the in/out/don't–know buttons for the six possible powders would be shown to the examinee.
- For a given task, if the examinee indicates that a solution is final, the task is determined to be completed.

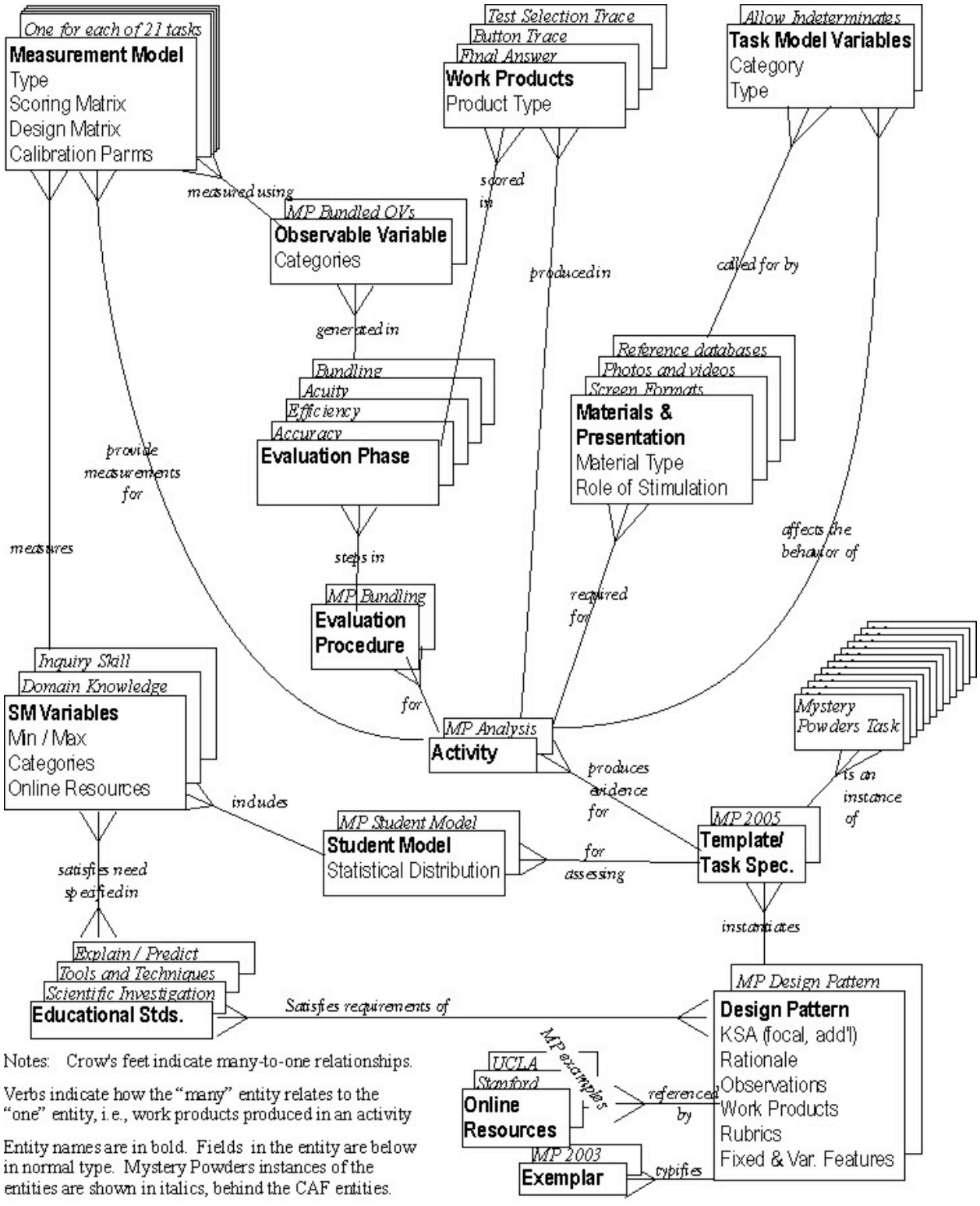
Finally, we used the spreadsheet to work out the way in which task performance would be evaluated. Most of the people who piloted the spreadsheet choose the optimal test—that is, the one that maximally reduces the solution space, most of the time, leading us to use dichotomous rather than a partial-credit scoring for evaluating the efficiency of the final solution. Because more than half the pilot–testers correctly identified the makeup of the average Mystery Powder, we also chose to measure accuracy using a dichotomous rather than a partial credit model. Based on our pilot data, partial credit made sense for evaluating the extent to which an examinee drew every possible conclusion from the available evidence at each step in solving the task.

5.0 Conceptual Assessment Framework for MP-QTI

5.1 Design Decisions

The structures in *conceptual assessment framework* layer of assessment design also reflect the assessment argument for Mystery Powders, but they move from the narrative form of *domain modeling* towards the details and machinery of operational assessments. In the *conceptual assessment framework* (CAF) we begin to articulate the assessment argument sketched in the *design pattern* (see Figure 5) in terms of the elements and processes that are needed to implement an assessment that embodies that argument. The structures in the CAF are expressed as objects such as variables, task schemas, and scoring mechanisms. The substance takes the form of particular values for these variables. Figure 6 shows the object structures and relationships in the CAF, as well as the particular values of those objects for the Mystery Powders example. This particular web of objects (the structure is shown in Figure 6), in which this technical level of design is expressed, is a PADI *template*. This section works through the *template*, with the discussion organized in terms of the roles the constituent elements play in terms of Student, Evidence, and Task Models.

Figure 6. Relationships of Models within the PADI CAF for Mystery Powders



One way to conceptualize the CAF is as machinery for generating assessment blueprints. Its structure coordinates the substantive, statistical, and operational aspects of an assessment. In the CAF, many design decisions are put into place to give concrete shape to the assessments we generate. These decisions include the kinds of statistical models that will be used, the materials that will characterize the examinee work environment, and the procedures that will be used to evaluate examinees' work. When we have done the work in the CAF layer, we will have the assessment argument expressed in operational terms, primed to generate a family of tasks and attendant processes that inform the target inference about examinee proficiency.

The CAF, as illustrated in Figure 6, is organized according to three models that correspond to the primary components of an assessment argument – the Student Model, Task Model, and Evidence Model. These models work in concert to provide the technical detail required for implementation such as specifications, operational requirements, statistical models, and details of rubrics. Claims, which in the *design pattern* were expressed as knowledge, skills, and abilities, are operationalized in terms of the variables in the Student Model. There can be one or several variables in an Student Model, and the Student Model can take a form as simple as an overall score across tasks and as complex as a multivariate item response theory or latent class model. The Student Model Variables link between examinees' performances on tasks and the claim(s) we wish to make about examinee proficiency. A probability distribution over these variables is used to express what one knows about an examinee at a given point in time (Mislevy, 1994).

The Task Model comprises the components necessary to lay out the features of the environment in which the examinee interacts with the task. Figure 6 shows the components of the Task Model: Task Model Variables, Materials and Presentation, and Work Products. This is where the Characteristics and Variable Features, as well as Potential Work Products from the *design pattern* will be represented in terms of stimulus materials. A variety of Potential Observations and Rubrics were identified in the *design pattern*, which linked Potential Work Products to KSAs. Each would have its own strengths and weaknesses, implementation costs and learning benefits. The designers choose the subset among these alternatives that best fit the purposes, resources, and context of the assessment being designed. These more specific forms are expressed in the Evidence Model. The Measurement Model, part of the Evidence Model, describes a mathematical function that relates observable evidence to Student Model Variables. The overview page of the *template* for MP-QTI is shown as Figure 7. We will now describe the particular design decisions made for the Student, Task, and Evidence Models for MP-QTI.

Figure 7. Mystery Powders Simulation Template

Mystery Powders - Simulation | Template 1996 [View Tree | Convert to Task Spec | Duplicate | Export | Delete]

Title:	[Edit]	Mystery Powders - Simulation
Summary	[Edit]	This template is for potential Mystery Powders assessments in a simulated environment, targeted to High School students. The student is presented with a mystery powder, or series of powders, the composition(s) of which the student must determine, given 6 potential powders in a mix: flour, cornstarch, salt (sodium chloride), sugar (sucrose), plaster (calcium hydroxide), and baking soda (sodium bicarbonate). The student selects from 6 chemical and physical tests that can be performed on a mixture, and is shown what observations are available from the test results. Vinegar and iodine are chemical tests; water, heat, taste, and visual inspection are physical tests.
Type	[Edit]	[View]
Student Model Summary	[Edit]	A bivariate model with variables for Domain Knowledge and Inquiry Skill.
Student Models	[Edit]	Mystery Powders SM. The student model includes two variables: 1) domain knowledge, a familiarity with the powder c...
Measurement Model Summary	[Edit]	Linking the observable variables to the estimation of SMVs is the Measurement Model, a conditional probability distribution for OVs given SMVs. The current version of the PADI Design System utilizes a specific psychometric model, the Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM; Adams, Wilson, & Wang, 1997), to describe the location of examinees' proficiency along one or more ability continua. For the Mystery Powders simulation, the OVs Accuracy, Acuity, and Efficiency are bundled into a final OV that is linked to estimations of the SMVs, Domain Knowledge and Inquiry Skill.
Evaluation Procedures Summary	[Edit]	The evaluation procedure yields three OVS: Acuity, Accuracy, and Efficiency. Accuracy is produced through a single evaluation phase; acuity and efficiency each require two evaluation phases. Finally, the three OVs are bundled into a final, single OV.
Work Product Summary	[Edit]	The work products are: (1) selection of and sequence of tests, (2) deductions made about presence/absence of particular powders after each test, and (3) final solution for a given mixture.
Task Model Variable Summary	[Edit]	Task Model Variables include the choice of an administrative mode, use of repeated tasks, selection of items (e.g., adaptively, pre-specified), the particular sets of potential powders and potential tests, specification of min and max numbers of powders for a given mixture, and the inclusion of indeterminate powders.
Template-level Task Model Variables	[Edit]	Administration Type. Task may be administered via computer or via paper and pencil. Use of multiple tasks/items. Are multiple tasks or items utilized in a given assessment? Adaptive selection of subsequent tasks/items. Given an assessment with multiple task/items, is adaptive selection used for subsequent tasks/items? Number and types of Mystery Powders. For a Mystery Powders assessment, the number of potential powders for any given mixture is finite an... Number and types of Mystery Powder tests. For a Mystery Powders assessment, the number of potential tests that can be applied to any given mix... Mystery Powders: Specify Min and Max?. A Mystery Powders assessment may or may not specify the mininimum and maximum numbers of powders for... Mystery Powders: include indeterminate results. Whether or not elements may be included in the Mystery Powder of which the presence or absence canno...
Task Model Variable Settings	[Edit]	[View]
Materials and Presentation Requirements	[Edit]	A computer with a browser and Internet access The particular presentation screens used The photos and videos used to show the results of a test (a file of photos and videos of all possible outcomes of the different tests)
Template-level Materials and Presentation	[Edit]	Computer with Browser and Internet Access. Mystery Powders reagent screen. Show the results of a reagent on the mystery powder. Allow student to declare that they have final a...
Materials and Presentation Settings	[Edit]	[View]
Activities Summary	[Edit]	The Activity listed below is generic, encompassing all 480 Mystery Powder items (62 combinations of powders in conjunction with all possible minimum and maximum settings from 1 to 5). A related Task Template provides settings and the associated activity for one particular Mystery Powders item. In this Activity, Work Products (interim and final solutions to a given powders, as well as selected tests) are converted to OVs using particular source file databases. The OVs, Accuracy, Acuity, and Efficiency, are combined into a final, bundled OV.
Activities	[Edit]	Mystery Powders Simulated Items. This activity represents all possible Mystery Powders items, given the potential powders, tests, and...
Tools for Examinee	[Edit]	
Exemplars	[Edit]	
Educational Standards	[Edit]	Unifying Concepts 1.3. Constancy, change, and measurement NSES 8ASI1.2. Design and conduct a scientific investigation. Students should develop general abilities, such as sy... NSES 8ASI1.3. Use appropriate tools and techniques to gather, analyze, and interpret data. The use of tools and te... NSES 8ASI2.2. Current scientific knowledge and understanding guide scientific investigations. Different methods, c...
Design Patterns	[Edit]	Hypothetico-Deductive Problem-Solving in a Finite Space. Students are presented with a problem of determining the state of an object or system, and methods f... Mystery Powders - Generalized. This design pattern is for potential Mystery Powders assessments targeted to middle- or high-school ... Mystery Powders Design Pattern. This design pattern is for the Mystery Powders Assessment. The computerized assessment presents a m...
I am a kind of	[Edit]	
These are kinds of me	[Edit]	Mystery Powders Sim - Sugar, Soda, & Salt 2/4. This task specification is for one specific Mystery Powders Simulated item: sugar, soda, and salt wi...
These are parts of me	[Edit]	
Online resources	[Edit]	
References	[Edit]	
I am a part of		

5.1.1 Student Model

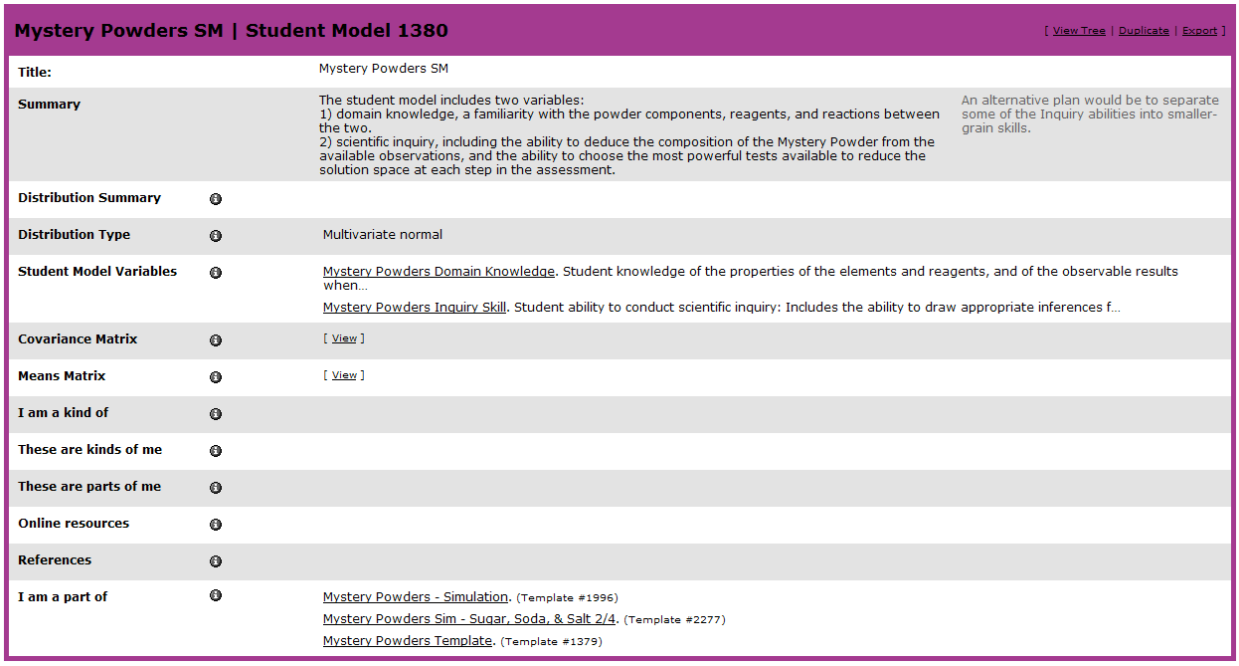
As mentioned previously with regard to the Mystery Powder *design pattern*, the claim of the assessment argument (see Figure 4) corresponds to the Student Model – the Knowledge, Skills, and Abilities we wish to assess. For Mystery Powders we noted that the Focal KSAs are hypothetical reasoning, deductive reasoning, and strategic efficiency. To instantiate the assessment argument at this CAF layer, it was necessary to define a Student Model and the Student Model Variables comprising this model. As indicated in Figure 6 (lower left), specific Student Model Variables serve as the link between examinees' performances on tasks (through the Measurement Model) and the claim(s) we wish to make about examinee proficiency (based on the Student Model). Upon considering the Focal KSAs outlined in the Mystery Powders *design pattern* (Figure 5), we decided that Mystery Powders-QTI will provide information about examinees in terms of two Student Model Variables (SMVs):

1. Domain Knowledge - understanding of the properties of the powders and the interpretation of the results of test using the available reagents and physical processes, and

2. Inquiry Skill – the ability to consider the entire space of hypothetical Mystery Powder combinations given the minimum and maximum numbers of powders in a mixture, the ability to determine which test would be likely to produce the most useful evidence under a given circumstance, and the ability to draw valid deductions from available evidence (test results).

These Student Model Variables were defined on the Mystery Powders Simulation (MP-QTI) *template,* the summary of which was provided in Figure 7. Figure 8 is a designer's view of the Student Model object itself in the template; and Figure 9 is a view of one of the Student Model Variables, Inquiry Skill.

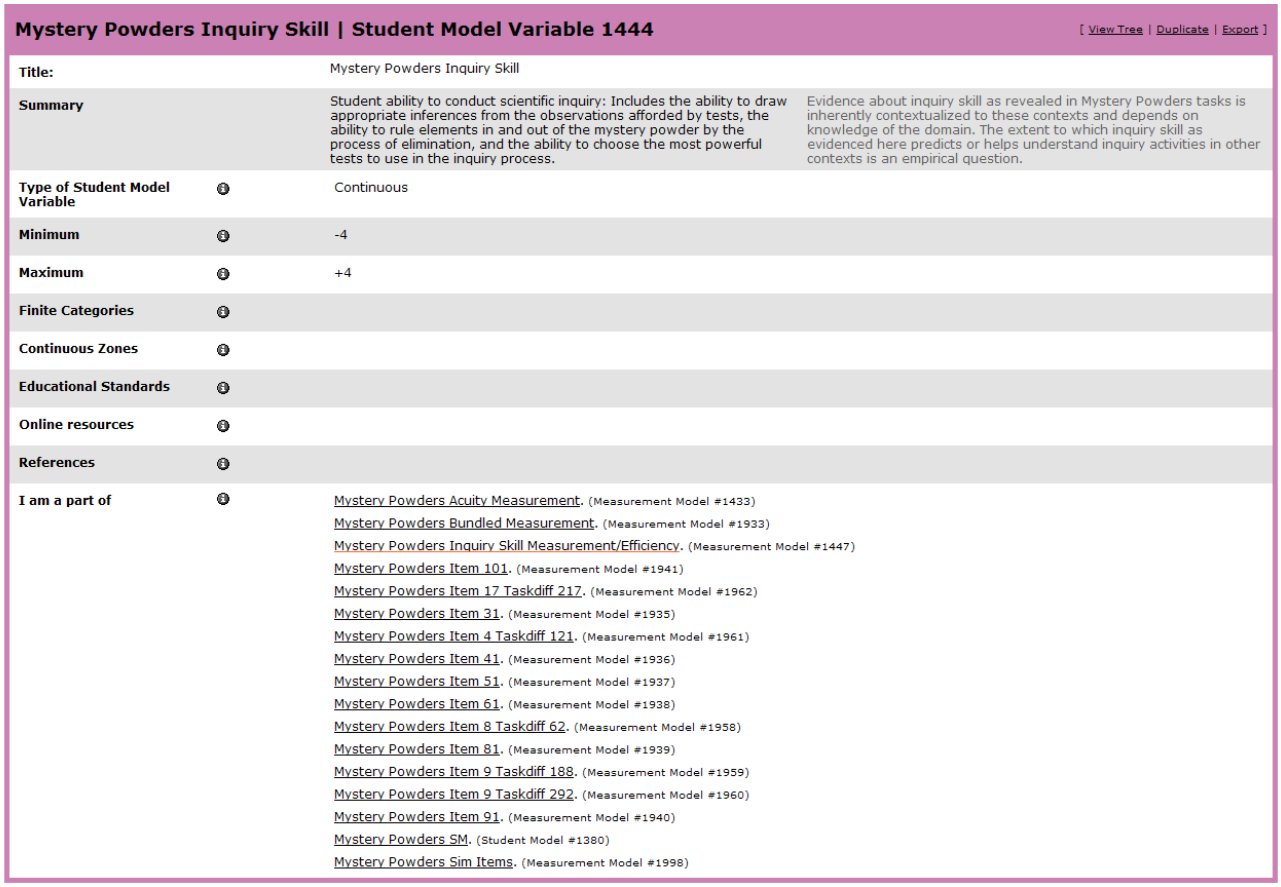
Figure 8. Student Model for Mystery Powders Simulation Template



Mystery Powders SM | Student Model 1380 [View Tree | Duplicate | Export]

Title:	Mystery Powders SM	
Summary	The student model includes two variables: 1) domain knowledge, a familiarity with the powder components, reagents, and reactions between the two. 2) scientific inquiry, including the ability to deduce the composition of the Mystery Powder from the available observations, and the ability to choose the most powerful tests available to reduce the solution space at each step in the assessment.	An alternative plan would be to separate some of the Inquiry abilities into smaller-grain skills.
Distribution Summary		
Distribution Type	Multivariate normal	
Student Model Variables	Mystery Powders Domain Knowledge. Student knowledge of the properties of the elements and reagents, and of the observable results when... Mystery Powders Inquiry Skill. Student ability to conduct scientific inquiry: Includes the ability to draw appropriate inferences f...	
Covariance Matrix	[View]	
Means Matrix	[View]	
I am a kind of		
These are kinds of me		
These are parts of me		
Online resources		
References		
I am a part of	Mystery Powders - Simulation. (Template #1996) Mystery Powders Sim - Sugar, Soda, & Salt 2/4. (Template #2277) Mystery Powders Template. (Template #1379)	

Figure 9. Inquiry Skill Student Model Variable for Mystery Powders Simulation



Mystery Powders Inquiry Skill | Student Model Variable 1444 [View Tree | Duplicate | Export]

Title:	Mystery Powders Inquiry Skill	
Summary	Student ability to conduct scientific inquiry: Includes the ability to draw appropriate inferences from the observations afforded by tests, the ability to rule elements in and out of the mystery powder by the process of elimination, and the ability to choose the most powerful tests to use in the inquiry process.	Evidence about inquiry skill as revealed in Mystery Powders tasks is inherently contextualized to these contexts and depends on knowledge of the domain. The extent to which inquiry skill as evidenced here predicts or helps understand inquiry activities in other contexts is an empirical question.
Type of Student Model Variable	Continuous	
Minimum	-4	
Maximum	+4	
Finite Categories		
Continuous Zones		
Educational Standards		
Online resources		
References		
I am a part of	Mystery Powders Acuity Measurement. (Measurement Model #1433) Mystery Powders Bundled Measurement. (Measurement Model #1933) Mystery Powders Inquiry Skill Measurement/Efficiency. (Measurement Model #1447) Mystery Powders Item 101. (Measurement Model #1941) Mystery Powders Item 17 Taskdiff 217. (Measurement Model #1962) Mystery Powders Item 31. (Measurement Model #1935) Mystery Powders Item 4 Taskdiff 121. (Measurement Model #1961) Mystery Powders Item 41. (Measurement Model #1936) Mystery Powders Item 51. (Measurement Model #1937) Mystery Powders Item 61. (Measurement Model #1938) Mystery Powders Item 8 Taskdiff 62. (Measurement Model #1958) Mystery Powders Item 81. (Measurement Model #1939) Mystery Powders Item 9 Taskdiff 188. (Measurement Model #1959) Mystery Powders Item 9 Taskdiff 292. (Measurement Model #1960) Mystery Powders Item 91. (Measurement Model #1940) Mystery Powders SM. (Student Model #1380) Mystery Powders Sim Items. (Measurement Model #1998)	

Other SMVs could have been entertained and, in fact, could be used with the same tasks in other contexts or for different purposes. Because an examinee might improve over the course of the assessment, a plausible SMV might be the rate of improvement. Another could be alacrity—that is, the examinee's capability to carry out investigations rapidly. Several considerations went into our choice of these Student Model Variables for the current demonstration. One is that these SMVs reflect key aspects of science learning consistent with national standards such as the National Science Education Standards (National Research Council, 1996). We can consider constructs of domain knowledge and inquiry skills to underlie many science assessments. Here we considered deductive reasoning, hypothetical reasoning, and strategic efficiency as inquiry skills involved in solving Mystery Powders for MP-QTI, but did not plan sufficiently detailed observations to sort them out—just their joint application as reflected by students' choices of tests and evaluations of their implications for whether powders were ruled in or out. We reconsidered the importance of domain knowledge – clearly a factor in determining the properties of powders in interaction with the various tests. We assumed that the ability to read instructions is not a major factor in test performance of students for whom MP-QTI might be used. We also assumed that most examinees are familiar enough with the Internet and browser interfaces that these skills would not be plausible explanations of poor performance on MP-QTI. (Note that these assumptions would be satisfied in practice either by specific knowledge about examinees, as in classroom assessment, or by offering sufficient instruction and practice, as in large-scale use.)

5.1.2 Task Model

The evidence to be gathered to support the claim of an assessment argument must be elicited through particular task structures, which are laid out in the Task Model. The Task Model is where the Characteristic and Variable Features, as well as Potential Work Products, of the *design pattern* will be more technically specified in terms of stimulus materials. In the Mystery Powders *design pattern* (Figure 5), we noted that Characteristic Features of a Mystery Powders assessment include a finite set of potential powders, a finite set of potential tests, known outcomes of the test, and known ways in which the presence of some powders may obscure evidence of the presence of other powders. In addition, tasks may vary according to the particular powders in a given mixture, the minimum and maximum numbers of powders communicated to the student, the presentation of prior results to earlier tasks, and the use of indeterminate powder combinations. We also noted Potential Work Products that include the selection of what test to perform, deductions made following each test, and the final conclusion with regard to the powder composition.

At the CAF layer, we flesh out the elements of the Task Model, characterizing the particular stimulus materials of an examinee's work environment. The PADI *template* provides an Activity structure that describes the structure of all potential tasks. In general, Activities constitute the major components of a *template* and are used to structure the generation, collection, and scoring of evidence. An Activity contains a group of related objects, including Materials and Presentation, Work Products, Evaluative Phases, Observable Variables, and Measurement Models. The Mystery Powders Simulation *template* contains one Activity—the presentation of a yet-to-be-specified mixture of

powders to be resolved—which, in combination with pre-specified settings for minimum and maximum numbers of powders, constitutes a task.

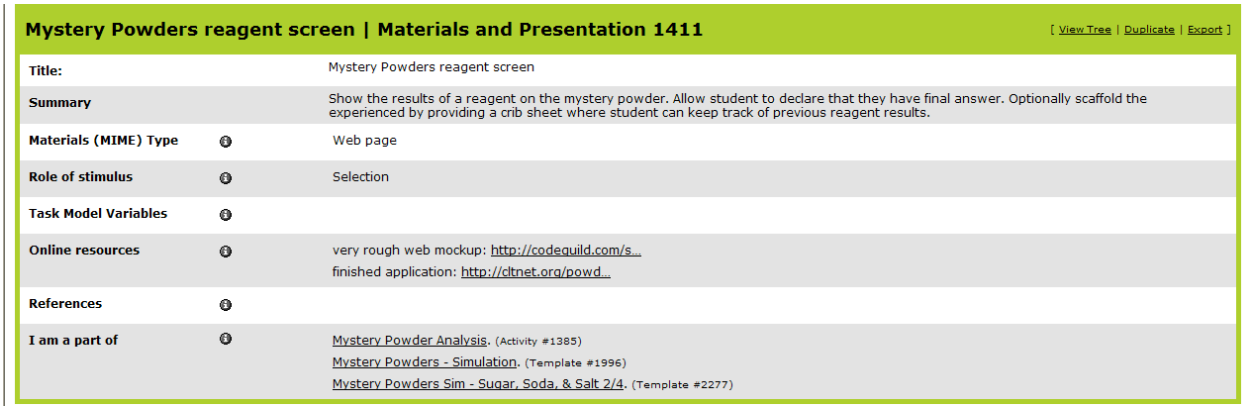
Figure 6 and the summary page of the *template* (Figure 7) provide the components of the Task Model—Task Model Variables, Materials and Presentation, and Work Products. In the middle of Figure 7 is a section for *template*-level Task Model Variables (TMVs). TMVs "describe key features of stimulus materials or relationships among them, tools and affordances (i.e., action possibilities) made available to examinees, or other aspects of the environments in which examinees work" (Mislevy & Riconscente, 2005, p. 14). TMVs can be defined at the *template* level (applying to the entire task) or at the Activity level (applying to materials or other conditions local to a particular activity within a multipart task). TMVs may be given specific settings (values) or may be left undefined. For MP–QTI, *template*-level TMVs were defined based on the Variable Features outlined in the *design pattern* (Figure 5); these TMVs were given specific settings (in parentheses):

1. The administration mode of the assessment (computer-delivered)
2. The administration of multiple tasks (yes)
3. The selection algorithm for subsequent tasks (adaptive)
4. The number and set of potential powders (6 powders: flour, cornstarch, salt, sugar, plaster, and baking soda)
5. The number and set of potential tests (6 tests: vinegar, iodine, water, heat, taste, and visual inspection)
6. The specification of minimum and maximum numbers of powders (yes)
7. The use of indeterminate powder combinations (yes)

Directly linked to the TMVs (e.g., Figure 6) are the Materials and Presentation and Work Products sections, which are subject to the control of TMVs. Materials and Presentation objects designate features of the assessment environment in which examinees will produce the evidence for the assessment argument (Mislevy & Riconscente, 2005). Such specific features are at a more technical and detailed level than the features addressed in the *design pattern*. Materials and Presentation objects reflect the type of task presented to the examinee; the features of these materials will be quite different for a computer-based assessment task (e.g., task presentation via a computer screen) than for a lab–based paper–and–pencil assessment task (e.g., task presentation via a test booklet, coupled with lab-materials).

For MP-QTI (Figure 7), the Materials and Presentation include a computer monitor with Internet access, the particular presentation screens used, and the photos and videos used to show the results of a test. As an example, Figure 10 shows the designer's view of one of the Presentation Materials, namely the Mystery Powders Simulation reagent screen.

Figure 10. Reagent Screen Presentation Material for Mystery Powders Simulation



Mystery Powders reagent screen | Materials and Presentation 1411 [View Tree | Duplicate | Export]

Title:	Mystery Powders reagent screen
Summary	Show the results of a reagent on the mystery powder. Allow student to declare that they have final answer. Optionally scaffold the experienced by providing a crib sheet where student can keep track of previous reagent results.
Materials (MIME) Type	Web page
Role of stimulus	Selection
Task Model Variables	
Online resources	very rough web mockup: http://codequild.com/s... finished application: http://dltnet.org/powd...
References	
I am a part of	Mystery Powder Analysis. (Activity #1385) Mystery Powders - Simulation. (Template #1996) Mystery Powders Sim - Sugar, Soda, & Salt 2/4. (Template #2277)

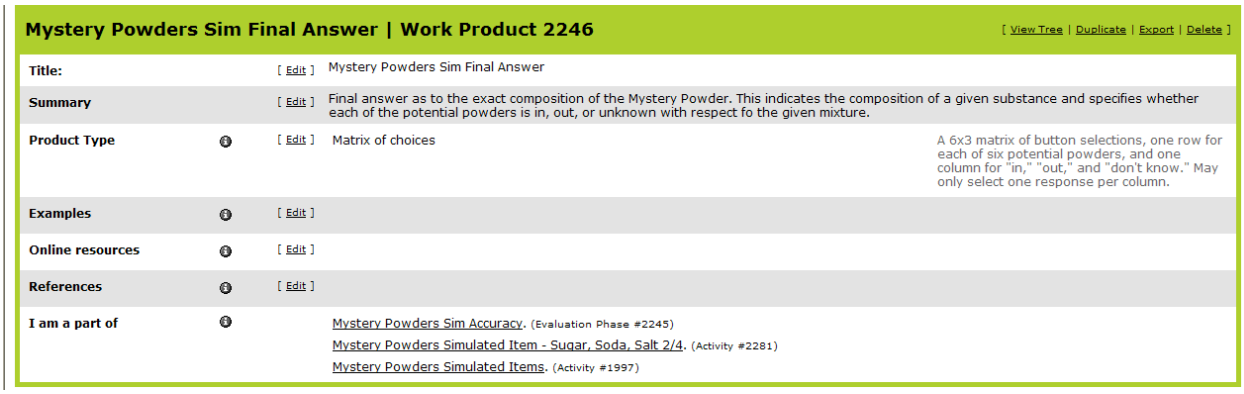
Work Products are also part of the Task Model. Work Products are the traces of examinees' responses to the given task—that which is evaluated to produce evidence. Depending on the nature of the task, Work Products could include written responses, oral responses, multiple-choice item responses, solution traces, or artifacts that the examinees build. In the Mystery Powders *design pattern* (Figure 5), we noted that Potential Work Products could include the selection of test to perform, settings for deductions that can be made at each step, and the conclusion with regard to the powder composition. The Work Products for Mystery Powders are records, known as traces, of the examinee's progress, such as which tests they chose, what they deduced at each step, and their final conclusion as to the composition of the powder. Specifically, after an examinee has conducted a test on her mixture of powders, she fills in a table of "In/Out/Don't Know" values for each powder. The matrices after each step are Work Products. Figure 11 shows the designer's view of a solution table Work Product in the PADI design system. We defined three Work Product objects for Mystery Powders Simulated Assessment (see Figure 7):

1. A trace of examinee's selection and sequence of experiments—for a given powder mixture accompanied by minimum and maximum settings, the examinee's sequence of selected experiments (e.g., taste 1st, heat 2nd, water 3rd, and iodine 4th),

2. A trace of examinee's deductions following each experiment—following each non-final experiment on a given powder mixture, 6-character strings (one string per experiment) representing the examinee's determinations for each of the 6 powders - in (1), out (0), or don't know (X), and

3. The final solution for a given mixture—for a given powder mixture, a 6-character string representing the examinee's final determination of whether each of the 6 powders is in (1), out (0), or cannot be determined (N).

Figure 11. Final Solution Table Work Product for Mystery Powders Simulation



The objects in the MP-QTI template that encompass the Task Model—namely TMVs, Materials and Presentation, and Work Products—characterize the examinee's work environment. Moving from the more narrative Mystery Powders *design pattern,* the Task Model portion of the MP-QTI *template* was specifically defined in the CAF layer to reflect a computer simulated assessment for Mystery Powders involving multiple tasks (different Mystery Powders) administered adaptively. A mixture is comprised of some combination of up to six powders, and six potential tests can be run to deduce which they are. The computer administration of the assessment traces the examinee's choices of tests and of both partial and final solutions.

5.1.3 Evidence Model

The evaluation of the data elicited through particular task structures, and the relationship of data to claim, fall within the scope of the Evidence Model. The Evidence Model addresses the question of which aspects of what student behavior(s) or performance(s) are hypothesized to bear on what variables in the Student Model. The Evidence Model includes the evaluative submodel and the statistical submodel. The evaluative submodel provides Evaluation Procedures, the rules for evaluating students' Work Products. This evaluative process results in Observable Variables. The statistical submodel, or the Measurement Model, provides a mathematical function that relates an Observable Variable to SMVs—specifically, a psychometric model that specifies probabilities for the values of Observable Variables conditional on values of Student Model Variables. The *design pattern* for Mystery Powders (Figure 5) provides Potential Work Products, Observations, and Rubrics for a range of potential assessments. For MP-QTI, Work Products, Observable Variables, Evaluation Procedures, and Measurement Models must be chosen to support the structure and goals of this computer-based assessment.

Background on Solving Mystery Powders tasks

In each step of solving a Mystery Powders task, the examinee chooses one of the previously unadministered tests and learns the results of applying that test to the mixture. The results of each test provide evidence that allows the examinee to rule in some potential powders as components of the mixture and rule out others—in Newell and Simon's (1972) terms, it reduces the problem space. A sequence of tests is necessary because only in rare instances can a single test provide sufficient evidence to support conclusions with regard to all six potential powders. At any step, the results of all the

tests carried out thus far together rule in or rule out a more complete collection of possibilities. At any step, each remaining test can potentially add more or less information than others. So at each step, there is a consideration of the optimality of the test the examinee chooses to carry out next.

The reduction of the solution space is an important aspect of solving Mystery Powders tasks and deserves some explanation. Table 4 shows the solution space reductions possible with the first test (within a given task). Consider the example of a minimum of 2 powders and a maximum of 5 powders. Initially, there would be 15 possible solutions with 2 powders each, 20 possible solutions with 3 powders each, 15 possible solutions with 4 powders each, and 6 possible solutions with 5 powders each—for a total of 56 possible solutions. It can be seen that if the taste test is given first and the resulting observation is sweet and salty, then the number of potential solutions is reduced to 15 because the deduction can be made that sugar and salt are present in the mixture.

Table 4. Solution Space Reductions Possible with the First Test

		Totals for each exact number of elements						Sum	Total/
Test	**Outcome**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**2-5**	**Average**
Taste	Tasteless	4	6	4	1	0	0	11	24.01
Taste	Sweet	1	4	6	4	1	0	15	36.97
Taste	Salty	1	4	6	4	1	0	15	36.97
Taste	Sweet and Salty	0	1	4	6	4	1	15	36.97
Taste Totals and Average		6	15	20	15	6	1		2.41
Heat	Nothing	3	3	1	0	0	0	4	5.05
Heat	Brown	2	7	9	5	1	0	22	61.90
Heat	Caramelize	1	5	10	10	5	1	30	92.88
Heat Totals and Average		6	15	20	15	6	1		2.85
Iodine	Not blue	5	10	10	5	1	0	26	77.11
Iodine	Blue	1	5	10	10	5	1	30	92.88
Iodine Totals and Average		6	15	20	15	6	1		3.04
Vinegar	No Fizz	5	10	10	5	1	0	26	77.11
Vinegar	Fizz	1	5	10	10	5	1	30	92.88
Vinegar Totals & Average		6	15	20	15	6	1		3.04
Water	Dissolve	3	3	1	0	0	0	4	5.05
Water	Gooey mess	1	3	3	1	0	0	7	12.40
Water	Lumpy stays muddy	1	4	6	4	1	0	15	36.97
Water	Lumpy hardens	1	5	10	10	5	1	30	92.88
Water Totals and Average		6	15	20	15	6	1		2.63
Visual	Crystal	2	1	0	0	0	0	1	0.00
Visual	Powder	4	6	4	1	0	0	11	24.01
Visual	Mixture	0	8	16	14	6	1	44	151.56
Visual Total		6	15	20	15	6	1		3.14

Mystery Powders is a deterministic assessment in that the universe of tasks and outcomes is sufficiently small and every situation can be enumerated. The domain is as follows:

- 62 possible combinations of between 1 and 5 of the 6 potential powders
- 480 tasks (powder combinations paired with each possible minimum and maximum)
- 24 unique observations from the six tests
- 1062 unique combinations of these 24 observations
- 4096 unique combinations of tests and powders, each of which results in 1 of the 1062 combinations of observations
- 4251 unique combinations of the observations, minimums, and maximums (associated with the deductions an examinee can make)
- 32,145 combinations of prior observations, minimums, maximums, and next test selected.

As discussed in the following sections, it is possible to evaluate every Observable Variable in the demonstration exactly in terms of the relationships between Work Products and Observational Variables.

5.1.4 The Evaluative Submodel

For the evaluative submodel for the adaptive, computer-delivered environment of the MP-QTI, an automated scoring algorithm was chosen. Not only would such an approach fit nicely with the need to select tasks adaptively, but it would allow us to show how the PADI *template* structures would be able to specify automated scoring requirements for a fully interactive real-time delivery system. A variety of evaluative Potential Rubrics were identified in the *design pattern*, which linked Potential Work Products to KSAs. Each may have its own strengths and weaknesses, costs and learning benefits. Choices among them and specific forms are now chosen to fit the purposes, resources, and context of this particular assessment. These more specific forms are detailed in the Evidence Model and again expressed in the form of objects in PADI *templates.*

For the Mystery Powders Simulation, evaluation procedures are used to evaluate, from Work Products, three Observable Variables:

- *Acuity:* While proceeding from each test to the next, how acutely has the examinee made every possible deduction from the observational evidence following each selected test?
- *Efficiency:* How close to optimal are the examinee's selections of each next test (e.g., does the choice optimally reduce the decision space?)
- *Accuracy:* Does the examinee correctly identify the mixture in the end?

MP-QTI uses three levels of evaluative phases: (1) stepwise phases (following each test of the mixture) for Acuity and Efficiency, (2) calculation of dichotomous OVs for Accuracy and Efficiency and a 4–level Final Acuity OV, then (3) a 16–level bundle of the three OVs. Other Observable Variables could have been defined, and different evaluation procedures from the ones described below could have been proposed to determine their values. No claim is made that the ones detailed below are optimal either instructionally or psychometrically; this could be the basis of an interesting study for improving MP-QTI for operational use in some particular context. Rather, these choices are meant only to be sufficiently plausible and consistent with pilot studies to be used for our main purpose, namely illustrating the use of the PADI design system structures.

The examinee's final solution for a given mixture (Work Product) is the basis of the Observable Variable Accuracy. Accuracy is a dichotomous (0-1) measure of whether the examinee produced the right answer, correctly indicating for each of the six possible powders whether it is in the mixture, out of the mixture, or that there cannot be enough information to tell (indeterminate). Guessing plays a role, and it is possible for the examinee to get the right answer without enough evidence to support the guess. This evaluative phase is as follows (see Appendix A for details, including the designer's view of the corresponding Evaluation Phase object in the MP-QTI PADI *template*):

- The examinee's final Work Product serves as input to the evaluation algorithm.
- For the given task, the Work Product string is compared to the "optimal" deductions for a given task, using an evaluation algorithm.
- The output OV *Accuracy* is created with 1 = right and 0 = wrong.

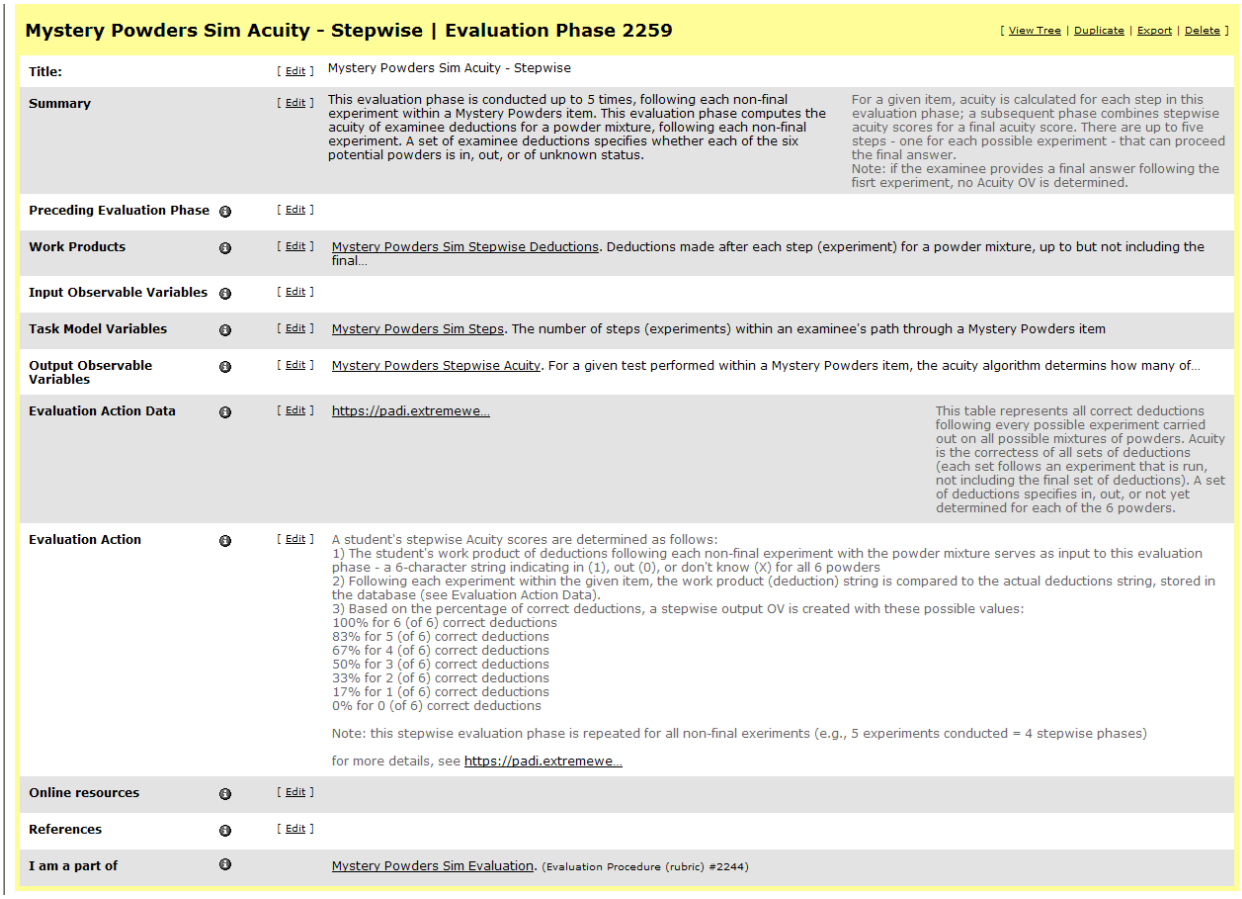
The OV Acuity is based on the trace of examinees' deductions following each test (all Work Products defined by solution matrices following each test). The Final Acuity OV is a partial credit (0-3) measure of how well the examinee draws appropriate inferences from the cumulative observations available at the end of a given test of the mixture. The evaluation of the examinee's deduction, resulting in the OV Acuity, is carried out in two phases: a stepwise phase and a final phase (see Appendix B for details). Since stepwise (and final) Acuity OVs are based on intermediate steps within a Mystery Powders task, if the examinee gives a final answer after only one experimental test, no Stepwise Acuity OVs are determined. Stepwise Acuity OVs are calculated individually and then summarized as a final Acuity OV. Stepwise Acuity OVs have six possible values, and are determined as follows:

- The input to the Stepwise Acuity Evaluation Phase is an examinee's Work Product of deductions following the *first* non-final experiment with the powder mixture—a 6–character string indicating in (1), out (0), or don't know (X) for all 6 powders.
- This Work Product (deduction) string is compared to the actual deductions string, stored in the database.
- Based on the percentage of correct deductions, a Step 1 Acuity OV is created with these possible values:
  100% for 6 (of 6) correct deductions

83% for 5 (of 6) correct deductions
67% for 4 (of 6) correct deductions
50% for 3 (of 6) correct deductions
33% for 2 (of 6) correct deductions
17% for 1 (of 6) correct deductions
0% for 0 (of 6) correct deductions

This procedure is repeated for all non-final steps. Figure 12 shows the corresponding Evaluation Phase from the designer's view—the same algorithm applied after the examinee indicates her deductions on the Solution Matrix after each experimental test. For example, if the examinee provided a final solution to the task after five experiments, this stepwise Evaluation Phase would be carried out four times, creating four stepwise OVs.

Figure 12. Stepwise Evaluation Phase for Acuity



Mystery Powders Sim Acuity - Stepwise | Evaluation Phase 2259

[View Tree | Duplicate | Export | Delete]

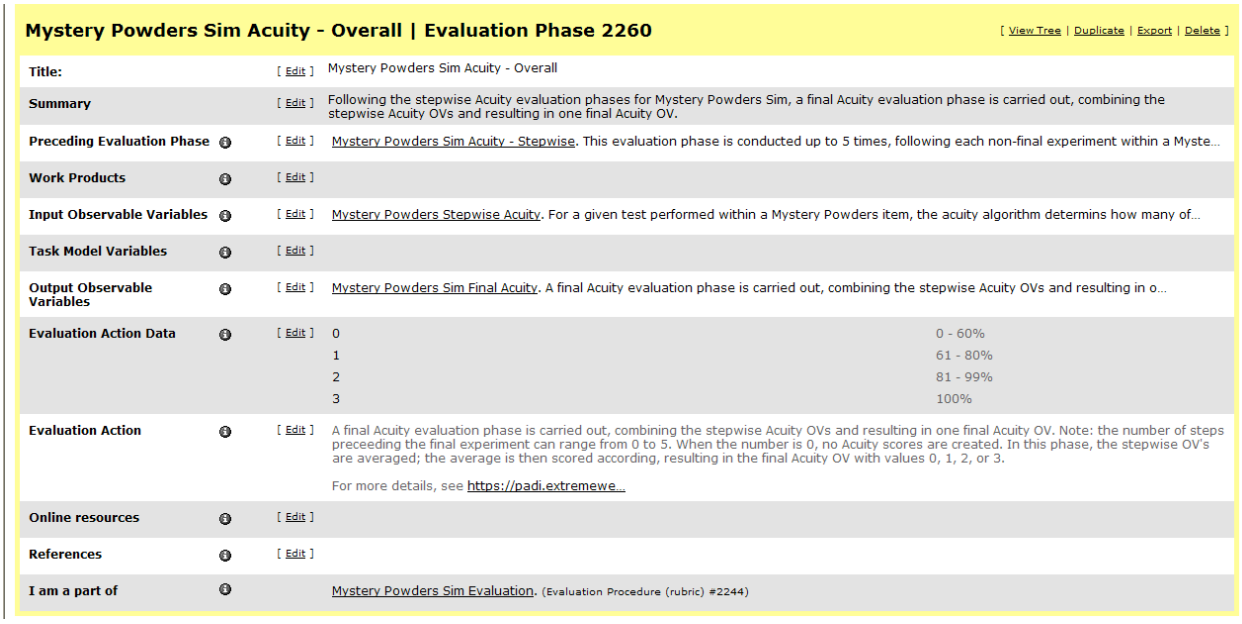
Title:	[Edit]	Mystery Powders Sim Acuity - Stepwise
Summary	[Edit]	This evaluation phase is conducted up to 5 times, following each non-final experiment within a Mystery Powders item. This evaluation phase computes the acuity of examinee deductions for a powder mixture, following each non-final experiment. A set of examinee deductions specifies whether each of the six potential powders is in, out, or of unknown status. For a given item, acuity is calculated for each step in this evaluation phase; a subsequent phase combines stepwise acuity scores for a final acuity score. There are up to five steps - one for each possible experiment - that can proceed the final answer. Note: if the examinee provides a final answer following the fisrt experiment, no Acuity OV is determined.
Preceding Evaluation Phase	[Edit]	
Work Products	[Edit]	Mystery Powders Sim Stepwise Deductions. Deductions made after each step (experiment) for a powder mixture, up to but not including the final...
Input Observable Variables	[Edit]	
Task Model Variables	[Edit]	Mystery Powders Sim Steps. The number of steps (experiments) within an examinee's path through a Mystery Powders item
Output Observable Variables	[Edit]	Mystery Powders Stepwise Acuity. For a given test performed within a Mystery Powders item, the acuity algorithm determins how many of...
Evaluation Action Data	[Edit]	https://padi.extremewe... This table represents all correct deductions following every possible experiment carried out on all possible mixtures of powders. Acuity is the correctess of all sets of deductions (each set follows an experiment that is run, not including the final set of deductions). A set of deductions specifies in, out, or not yet determined for each of the 6 powders.
Evaluation Action	[Edit]	A student's stepwise Acuity scores are determined as follows: 1) The student's work product of deductions following each non-final experiment with the powder mixture serves as input to this evaluation phase - a 6-character string indicating in (1), out (0), or don't know (X) for all 6 powders 2) Following each experiment within the given item, the work product (deduction) string is compared to the actual deductions string, stored in the database (see Evaluation Action Data). 3) Based on the percentage of correct deductions, a stepwise output OV is created with these possible values: 100% for 6 (of 6) correct deductions 83% for 5 (of 6) correct deductions 67% for 4 (of 6) correct deductions 50% for 3 (of 6) correct deductions 33% for 2 (of 6) correct deductions 17% for 1 (of 6) correct deductions 0% for 0 (of 6) correct deductions Note: this stepwise evaluation phase is repeated for all non-final exeriments (e.g., 5 experiments conducted = 4 stepwise phases) for more details, see https://padi.extremewe...
Online resources	[Edit]	
References	[Edit]	
I am a part of		Mystery Powders Sim Evaluation. (Evaluation Procedure (rubric) #2244)

A final Acuity Evaluation Phase is carried out, combining the stepwise Acuity OVs and resulting in one final Acuity OV. In this phase, the Stepwise Acuity OVs are averaged. The average is then categorized according to the following percentage ranges:

Score	Percentage Range (Average)
0	Up to 60%
1	61% - 80%
2	81% - 99%
3	100%

Thus, one final Acuity OV is created with a value of 0, 1, 2, or 3. (Note that the Stepwise Acuity OVs all have six values while the Final Acuity OV has summarized the average value in terms of just four categories.) The corresponding evaluation phase is shown from the designer's view in Figure 13.

Figure 13. Evaluation Phase for Final Acuity Observable Variable



A trace of examinees' selection and sequence of steps (Work Product) is the input for determining the OV called Efficiency. Efficiency is a dichotomous (0-1) measure of whether or not the examinee always chose the most efficient test at each step in the investigation. As with Acuity, stepwise Efficiency OVs are calculated individually and then combined into one final Efficiency OV (see Appendix C for details). The first stepwise Efficiency OVs is determined as follows:

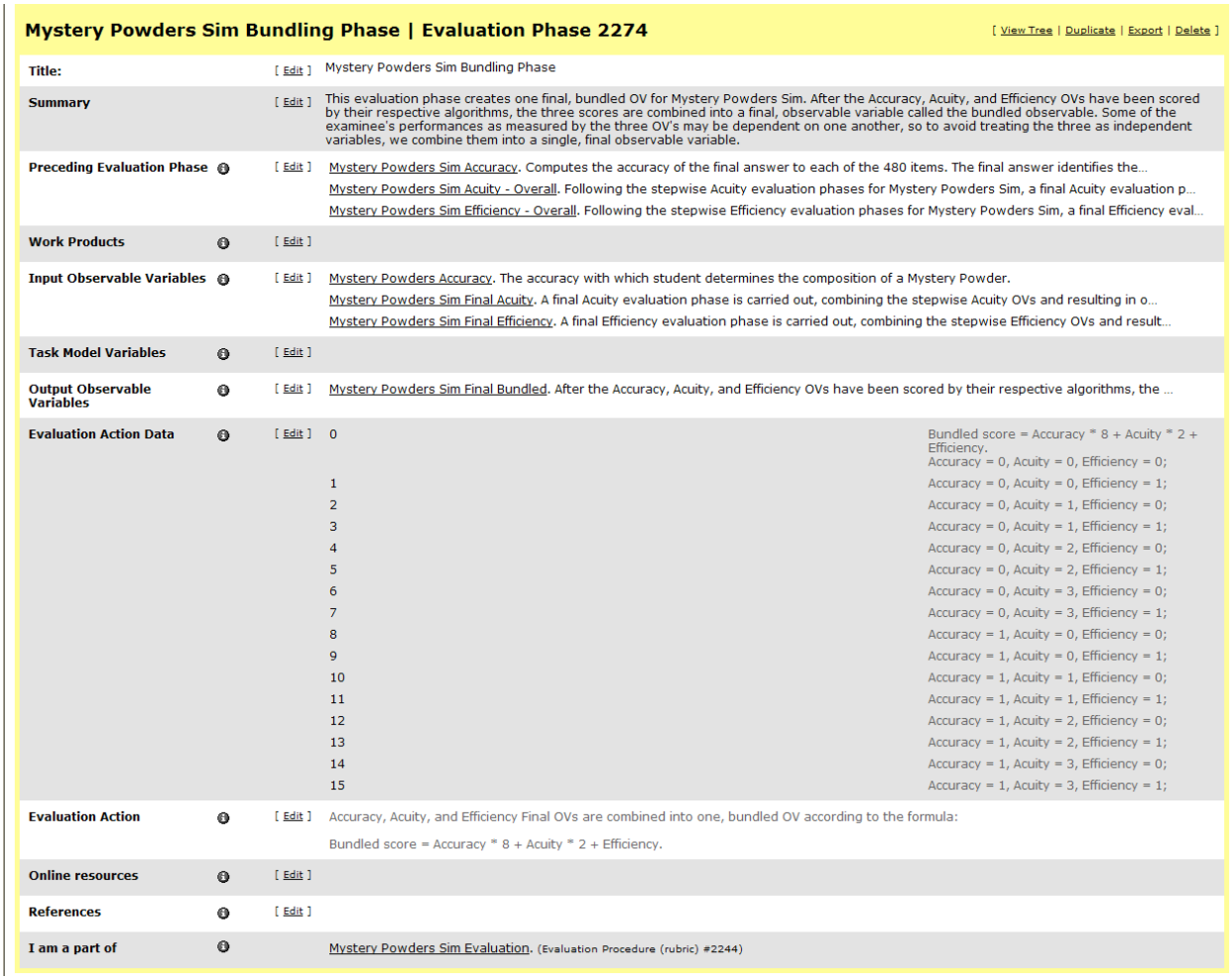
- The input to the stepwise Efficiency Evaluation Phase is an examinee's first choice of experiment, given a specified minimum and maximum numbers of powders.
- This first choice of experiment is compared to the optimal choice of experiment, retrieved from the appropriate Evaluation Data database.

- If the first choice of experiment is optimal (a value of 1.00), the Step 1 Efficiency OV is created with a score of 1. Otherwise it is given a score of 0.

This procedure is repeated for all steps within the task. Subsequent steps take previous experiments' choices and resulting observations into account. Stepwise OVs are created in this Evaluation Phase, corresponding to each choice of experimental test. A final Efficiency Evaluation Phase is carried out, combining the stepwise Efficiency OVs and resulting in one final Efficiency OV. If all of the stepwise OVs have a value of 1, the final Efficiency OV is assigned a value of 1; otherwise, the final Efficiency OV is assigned a value of 0. Thus, one final Efficiency OV is created, with a value of 0 or 1.

A final Evaluation Phase creates one bundled OV for Mystery Powders Simulation assessment (see Figure 14 for the designer's view of the corresponding Evaluation Phase). After the Accuracy, Acuity, and Efficiency OVs have been scored by their respective algorithms, the three scores are combined into a final Observable Variable called the Bundled Observable. Since some of the examinee's performances as measured by the three OV's may be dependent on one another, we avoid treating the three as conditionally independent variables by combining them into a single, final Observable Variable (Wilson & Adams, 1995). Exactly how this is accomplished is detailed in the following section. The final bundled OV is a combination of Accuracy, Acuity, and Efficiency and can assume all discrete values from 0 to 15.

Figure 14. Evaluation Phase for the Bundled Observable



Mystery Powders Sim Bundling Phase | Evaluation Phase 2274 [View Tree | Duplicate | Export | Delete]

Title:	[Edit]	Mystery Powders Sim Bundling Phase
Summary	[Edit]	This evaluation phase creates one final, bundled OV for Mystery Powders Sim. After the Accuracy, Acuity, and Efficiency OVs have been scored by their respective algorithms, the three scores are combined into a final, observable variable called the bundled observable. Some of the examinee's performances as measured by the three OV's may be dependent on one another, so to avoid treating the three as independent variables, we combine them into a single, final observable variable.
Preceding Evaluation Phase	[Edit]	Mystery Powders Sim Accuracy. Computes the accuracy of the final answer to each of the 480 items. The final answer identifies the... Mystery Powders Sim Acuity - Overall. Following the stepwise Acuity evaluation phases for Mystery Powders Sim, a final Acuity evaluation p... Mystery Powders Sim Efficiency - Overall. Following the stepwise Efficiency evaluation phases for Mystery Powders Sim, a final Efficiency eval...
Work Products	[Edit]	
Input Observable Variables	[Edit]	Mystery Powders Accuracy. The accuracy with which student determines the composition of a Mystery Powder. Mystery Powders Sim Final Acuity. A final Acuity evaluation phase is carried out, combining the stepwise Acuity OVs and resulting in o... Mystery Powders Sim Final Efficiency. A final Efficiency evaluation phase is carried out, combining the stepwise Efficiency OVs and result...
Task Model Variables	[Edit]	
Output Observable Variables	[Edit]	Mystery Powders Sim Final Bundled. After the Accuracy, Acuity, and Efficiency OVs have been scored by their respective algorithms, the ...
Evaluation Action Data	[Edit]	0 — Bundled score = Accuracy * 8 + Acuity * 2 + Efficiency. Accuracy = 0, Acuity = 0, Efficiency = 0; 1 — Accuracy = 0, Acuity = 0, Efficiency = 1; 2 — Accuracy = 0, Acuity = 1, Efficiency = 0; 3 — Accuracy = 0, Acuity = 1, Efficiency = 1; 4 — Accuracy = 0, Acuity = 2, Efficiency = 0; 5 — Accuracy = 0, Acuity = 2, Efficiency = 1; 6 — Accuracy = 0, Acuity = 3, Efficiency = 0; 7 — Accuracy = 0, Acuity = 3, Efficiency = 1; 8 — Accuracy = 1, Acuity = 0, Efficiency = 0; 9 — Accuracy = 1, Acuity = 0, Efficiency = 1; 10 — Accuracy = 1, Acuity = 1, Efficiency = 0; 11 — Accuracy = 1, Acuity = 1, Efficiency = 1; 12 — Accuracy = 1, Acuity = 2, Efficiency = 0; 13 — Accuracy = 1, Acuity = 2, Efficiency = 1; 14 — Accuracy = 1, Acuity = 3, Efficiency = 0; 15 — Accuracy = 1, Acuity = 3, Efficiency = 1;
Evaluation Action	[Edit]	Accuracy, Acuity, and Efficiency Final OVs are combined into one, bundled OV according to the formula: Bundled score = Accuracy * 8 + Acuity * 2 + Efficiency.
Online resources	[Edit]	
References	[Edit]	
I am a part of		Mystery Powders Sim Evaluation. (Evaluation Procedure (rubric) #2244)

5.1.5 The Statistical Submodel

In order to ground an inference about how much an examinee knows or what an examinee can do, the Observable Variables are interpreted as evidence about the examinees' unobservable Knowledge, Skills, or Abilities (KSAs). Examinee proficiency is cast as settings for one or more Student Model Variables that characterize KSAs in a perspective and at a grain size (granularity) that suits the purpose of the assessment. The Student Model contains SMVs and a probability distribution that represents the analyst's knowledge about their values at a given point in time.

Linking the OVs to the estimation of SMVs is the Measurement Model, a conditional probability distribution for OVs at each possible value of the SMVs. The current version of the PADI design system utilizes a specific psychometric model, the Multidimensional Random Coefficient Multinomial Logit Model (MRCMLM; Adams, Wilson, & Wang, 1997) to describe the location of examinees' proficiencies along one or more ability continua. In a simple case where only a single facet of knowledge is being assessed, there will only be a single SMV. However, in science assessments, it can be useful to interpret performance in terms of two or more facets of knowledge or ability. In these cases a

multivariate Student Model with a collection of SMVs and a multivariate probability distribution is used to express what is known about a examinee's ability (Mislevy & Riconscente, 2005).

To illustrate aspects of the PADI design framework, we chose to use a multivariate measurement model. MP-QTI is designed to measure the two examinee KSAs previously described as part of the Student Model: domain knowledge and inquiry skill. Recall that in the MP-QTI context, domain knowledge addresses the examinee's knowledge about how powders and combinations of powders behave in the presence of the reagents, and inquiry skill is the ability to make the available inferences from the evidence at hand and to choose the most powerful among the available diagnostic tests.

The statistical submodel, operationalized in PADI as Measurement Models, is a part of the Evidence Model that describes a mathematical function relating observable evidence to Student Model Variables. The functional form is a special case of MRCMLM that accommodates more complex measurement tasks involving multidimensionality, partial credit, rating scales, and conditional independence. We posited the following dependence relationships:

- The Accuracy OV depends on the domain knowledge SMV only.
- The Acuity OV depends on both the domain knowledge and inquiry skill SMVs.
- The Efficiency OV depends on the inquiry skill SMV only.

As noted in the previous section, conditional dependence among the three observables from each given task led us to model a bundled Observable Variable that affects both the relationships listed above and the interrelationships of observables within a task beyond the relationships caused by their dependence on the SMVs. Suppose Task j has *Kj* possible responses, indexed by *k*. The mathematical form for the conditional probability of the particular response k^* to Task *j* is given as follows:

$$P\left(X_j = k^* \middle| \theta, \eta_j, A_j, B_j\right) = \frac{\exp\left(b_{jk^*}{}' \theta + a_{jk^*}{}' \eta_j\right)}{\sum_k \exp\left(b_{jk}{}' \theta + a_{jk}{}' \eta_j\right)},$$

where

$\theta = (\theta_{DK}, \theta_{IS})$ with *qDK* the SMV for domain knowledge and *qIS* the SMV for inquiry skill,

hj is a vector of difficulty parameters for Task *j*,

Aj, the MRCMLM "design matrix," contains a row *ajk* for each response category *k* that relates difficulty parameters to response categories,

Bj, the MRCMLM "scoring matrix," contains a row *bjk* for each response category indicating which elements of *q* it depends on.

The details of the MRCMLM are beyond the scope of this paper. What is important for present purposes is illustrating the use of PADI design objects to express a multivariate student model, with different Observable Variables depending on different combinations of Student Model Variables with conditional dependence of observables from the same task being handled by bundling, and with the same structure being used for all tasks in the domain although they may differ through their task parameters *hj*.

Figures 15-17 show the Measurement Model used in MP-QTI from the designer's view. The same structure is used for all possible tasks. Each task may have different difficulty parameters reflecting the fact that some tasks are more challenging than others based on the numbers and combinations of powders in their mixture and the maximum and minimums communicated to the examinee—all Task Model Variables in the MP-QTI *template*. In particular,

- The more powders that are in a mixture, the harder the task is to solve.
- The identity of the individual powders in the mixture affects the difficulty of the task—sugar is easiest to detect, flour hardest.
- The less the solution space is constrained by the given minimum and maximum number of powders an examinee is told might be in a mixture, the harder the task is to solve.
- If the mixture is indeterminate (i.e., some possibilities cannot be distinguished because the results from one test mask the results of subsequent tests), the task is harder.

Figure 15. Measurement Model for Mystery Powders Simulation

Mystery Powders Sim Items | Measurement Model 1998 [View Tree | Duplicate | Export | Delete]

Title:	[Edit]	Mystery Powders Sim Items	
Summary	[Edit]	Measurement model for all simulated Mystery Powders items, given potential powders, tests, and min and max settings	Accuracy is modeled as depending only on the Domain Knowledge SMV. Acuity is modeled as depending equally on Domain Knowledge and Inquiry Skill. Efficiency is modeled as depending only on Inquiry Skill
Type of Measurement Model	[Edit]	Item bundle	
Observable Variable	[Edit]	Mystery Powders Bundled OV. Combine Accuracy, Acuity and Efficiency into one Observed Variable.	
Student Model Variables	[Edit]	Mystery Powders Domain Knowledge. Student knowledge of the properties of the elements and reagents, and of the observable results when... Mystery Powders Inquiry Skill. Student ability to conduct scientific inquiry: Includes the ability to draw appropriate inferences f...	
Scoring Matrix	[Edit]	[View]	
Design Matrix	[Edit]	[View]	
Calibration Parameters	[Edit]	[View]	
Online resources	[Edit]		
References	[Edit]		
I am a part of		Mystery Powders Simulated Item - Sugar, Soda, Salt 2/4. (Activity #2281) Mystery Powders Simulated Items. (Activity #1997)	

Figure 16. Scoring Matrix

View Scoring Matrix

View Scoring Matrix that is part of Mystery Powders Sim Items.

[Edit Matrix]

Categories for OV: Mystery Powders Bundled OV	Values for SMV: Mystery Powders Domain Knowledge	Values for SMV: Mystery Powders Inquiry Skill
0	0	0
1	0	3
2	1	1
3	1	4
4	2	2
5	2	5
6	3	3
7	3	6
8	3	0
9	3	3
10	4	1
11	4	4
12	5	2
13	5	5
14	6	3
15	6	6

Comment:

List of Examples:

Figure 17. Design Matrix

View Design Matrix

View the Design Matrix that is part of Mystery Powders Sim Items.

[Edit Matrix]

Categories for OV: Mystery Powders Bundled OV	Difficulty of step 1	Difficulty of step 2	Difficulty of step 3	Difficulty of step 4	Difficulty of step 5	Difficulty of step 6	Difficulty of step 7	Difficulty of step 8	Difficulty of step 9	Difficulty of step 10	Difficulty of step 11	Difficulty of step 12	Difficulty of step 13	Difficulty of step 14	Difficulty of step 15
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Comment:

List of Examples:

The demonstration version of MP-QTI includes scoring with the BEAR Scoring Engine. The item parameters used in the demonstration are based on simulated data discussed in the next section.

6.0 Assessment Implementation

The next layer in the ECD assessment design scheme is *assessment implementation*. Implementation concerns creating the assessment pieces that the CAF structures specify. Having carried forth assessment design thus far with adherence to a shared argument, the designer is now poised to create multiple instances of tasks that may vary in their surface features but share the same underlying rationale and assessment argument. Design decisions are finalized in this layer. This section describes task authoring and assessment assembly, task calibration with simulated data, scoring and delivery algorithms, and presentation decisions for the MP-QTI assessment. All of these elements and algorithms are then used in the actual delivery and scoring of the assessment, the layer that will be described subsequently.

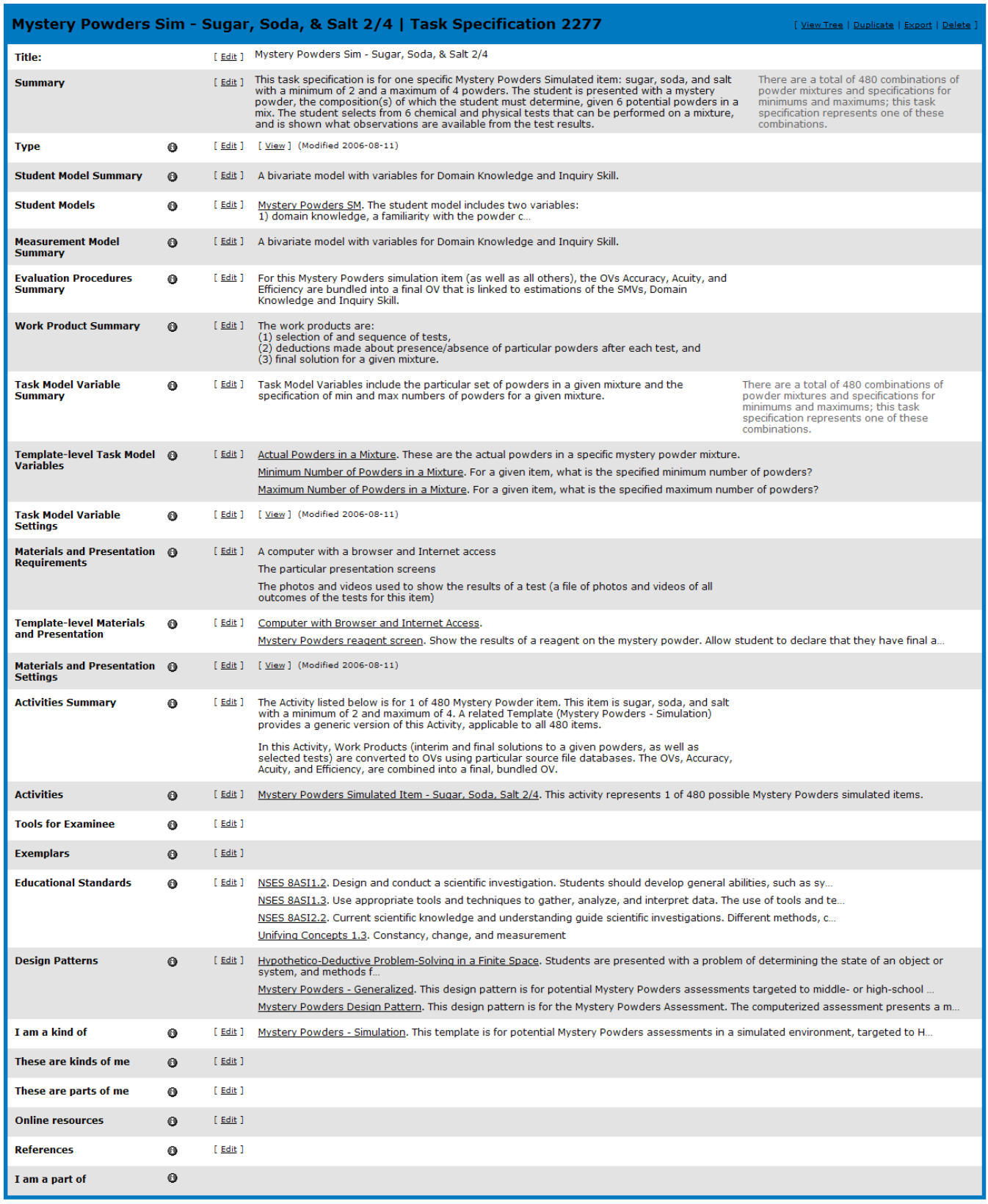
6.1 Task Authoring

The *template* for the Mystery Powders Simulation assessment (Figure 7) is a blueprint for all potential Mystery Powders tasks, yet it does not set forth the specifics for the universe of 480 MP-QTI tasks. An important step towards authoring an individual task involves the use of a *task specification*, a fully-specified variant of a *template* in the PADI design system. While *templates* are capable of generating families of tasks*, task specifications* are final plans for individual tasks. When a *task specification* is created from a *template*, the values of some attributes will be selected from among pre-specified options, while the values of other attributes will remain unchanged.

Figure 18 provides an overview of a *task specification* for one of the 480 potential Mystery Powders Simulation tasks. Similar *task specifications* could be authored for the remaining tasks. The task represented in Figure 18 is for the powder containing sugar, soda, and salt, with a minimum of 2 and a maximum of 4. This *task specification* uses the Student and Measurement Models found in the *template* (Figure 7), but the Task Model and evaluative submodel are more specific:

- The TMVs in the *template* are variables for the actual powders in the mixture, the minimum numbers of powders, and the maximum number of powders; the TMVs in this *task specification* are set to the values of sugar/soda/salt, 2, and 4.
- The Evaluation Procedures are the same as in the *template*, but given specific settings. For example, instead of utilizing the entire ‘Deductions’ database to evaluate the deductions made after each test within this task, only rows involving possible observations from this mixture (sugar, soda, salt) will be used (e.g., the entry '310004' is needed (visual mixture, dissolves in water, tastes sweet & salty), but '100001' is not (crystal & tasteless).
- Presentation materials are more specific. Web addresses or file names (URLs) are provided for graphics files that present photographs or video clips of the results of the analytic tests when applied to the mixture in this task. These Presentation Materials appear in many tasks because they are driven by the results of applying tests to mixtures that contain or do not contain particular powders. We will discuss these presentation materials in a separate section shortly.

Figure 18. Task Specification for Sugar, Soda, & Salt 2/4 #2277



Mystery Powders Sim - Sugar, Soda, & Salt 2/4 | Task Specification 2277

[View Tree | Duplicate | Export | Delete]

Title:	[Edit]	Mystery Powders Sim - Sugar, Soda, & Salt 2/4	
Summary	[Edit]	This task specification is for one specific Mystery Powders Simulated item: sugar, soda, and salt with a minimum of 2 and a maximum of 4 powders. The student is presented with a mystery powder, the composition(s) of which the student must determine, given 6 potential powders in a mix. The student selects from 6 chemical and physical tests that can be performed on a mixture, and is shown what observations are available from the test results.	There are a total of 480 combinations of powder mixtures and specifications for minimums and maximums; this task specification represents one of these combinations.
Type	[Edit]	[View] (Modified 2006-08-11)	
Student Model Summary	[Edit]	A bivariate model with variables for Domain Knowledge and Inquiry Skill.	
Student Models	[Edit]	Mystery Powders SM. The student model includes two variables: 1) domain knowledge, a familiarity with the powder c...	
Measurement Model Summary	[Edit]	A bivariate model with variables for Domain Knowledge and Inquiry Skill.	
Evaluation Procedures Summary	[Edit]	For this Mystery Powders simulation item (as well as all others), the OVs Accuracy, Acuity, and Efficiency are bundled into a final OV that is linked to estimations of the SMVs, Domain Knowledge and Inquiry Skill.	
Work Product Summary	[Edit]	The work products are: (1) selection of and sequence of tests, (2) deductions made about presence/absence of particular powders after each test, and (3) final solution for a given mixture.	
Task Model Variable Summary	[Edit]	Task Model Variables include the particular set of powders in a given mixture and the specification of min and max numbers of powders for a given mixture.	There are a total of 480 combinations of powder mixtures and specifications for minimums and maximums; this task specification represents one of these combinations.
Template-level Task Model Variables	[Edit]	Actual Powders in a Mixture. These are the actual powders in a specific mystery powder mixture. Minimum Number of Powders in a Mixture. For a given item, what is the specified minimum number of powders? Maximum Number of Powders in a Mixture. For a given item, what is the specified maximum number of powders?	
Task Model Variable Settings	[Edit]	[View] (Modified 2006-08-11)	
Materials and Presentation Requirements	[Edit]	A computer with a browser and Internet access The particular presentation screens The photos and videos used to show the results of a test (a file of photos and videos of all outcomes of the tests for this item)	
Template-level Materials and Presentation	[Edit]	Computer with Browser and Internet Access. Mystery Powders reagent screen. Show the results of a reagent on the mystery powder. Allow student to declare that they have final a...	
Materials and Presentation Settings	[Edit]	[View] (Modified 2006-08-11)	
Activities Summary	[Edit]	The Activity listed below is for 1 of 480 Mystery Powder item. This item is sugar, soda, and salt with a minimum of 2 and maximum of 4. A related Template (Mystery Powders - Simulation) provides a generic version of this Activity, applicable to all 480 items. In this Activity, Work Products (interim and final solutions to a given powders, as well as selected tests) are converted to OVs using particular source file databases. The OVs, Accuracy, Acuity, and Efficiency, are combined into a final, bundled OV.	
Activities	[Edit]	Mystery Powders Simulated Item - Sugar, Soda, Salt 2/4. This activity represents 1 of 480 possible Mystery Powders simulated items.	
Tools for Examinee	[Edit]		
Exemplars	[Edit]		
Educational Standards	[Edit]	NSES 8ASI1.2. Design and conduct a scientific investigation. Students should develop general abilities, such as sy... NSES 8ASI1.3. Use appropriate tools and techniques to gather, analyze, and interpret data. The use of tools and te... NSES 8ASI2.2. Current scientific knowledge and understanding guide scientific investigations. Different methods, c... Unifying Concepts 1.3. Constancy, change, and measurement	
Design Patterns	[Edit]	Hypothetico-Deductive Problem-Solving in a Finite Space. Students are presented with a problem of determining the state of an object or system, and methods f... Mystery Powders - Generalized. This design pattern is for potential Mystery Powders assessments targeted to middle- or high-school ... Mystery Powders Design Pattern. This design pattern is for the Mystery Powders Assessment. The computerized assessment presents a m...	
I am a kind of	[Edit]	Mystery Powders - Simulation. This template is for potential Mystery Powders assessments in a simulated environment, targeted to H...	
These are kinds of me	[Edit]		
These are parts of me	[Edit]		
Online resources	[Edit]		
References	[Edit]		
I am a part of			

6.2 Task Calibration

In practice, the difficulty parameters *hj* of the MRCML model used for test-level scoring would be estimated from the responses of students. Item and test fit statistics would be examined; alternative models might be estimated and compared. Different approaches to defining Observable Variables and comparisons of Measurement Models including and ignoring conditional dependence would be examined. For the purposes of the MP-QTI example, which is meant as a demonstration of the use of the design structures, the Measurement Model parameters are based on simulated data generated in accordance with the hypothesized difficulty structure, using the ConQuest computer program (Wu, Adams, & Wilson, 1998). The generation of simulated data used in the calibration is described in Appendix D.

6.3 Assembly Model

A valid assessment needs to include the right balance of tasks to cover all the domains and psychometric properties of interest, with enough tasks providing evidence on each of those dimensions to support inferences that are valid within predefined limits of statistical error. The logic for marshalling multiple tasks into an assessment is coordinated by the Assembly Model. This logic governs the operations of the *activity selection process* when an assessment is actually being delivered. The Assembly Model is a part of the evidence-centered design framework described in Mislevy, Steinberg, and Almond (2002), although it is not encompassed in the PADI object model.

As mentioned previously, we administer multiple powder mixtures as individual tasks. Task selection for the MP-QTI demonstration is based on Lord's flexilevel adaptive testing algorithm (1971, 1980). In flexilevel testing, a fixed set of items, twice as long as the intended test length, is arranged from easy to hard. Each examinee starts with the item in the middle. After every correct response, she takes the first more difficult item that she has not yet been presented. After every incorrect response, she takes the first item in the easier direction she has not yet taken. Each examinee ends up responding to a selected half of the total items adapted to her performance. An examinee who answers every item correctly takes the hard half of the items, an examinee who misses every item takes the easy half, and most examinees take some contiguous set of items somewhere in the middle.

For an MP-QTI demonstration assessment, a set of twenty-one tasks was selected that spanned a range from easy to hard tasks. This selection was based on the number of powders in the mixture (more is harder), and maximums and minimums (having the maximum and minimum number of powders further apart makes a task harder). Getting a satisfactory solution to the powders in the mixture constitutes a correct answer (either a correct final solution or mostly correct deductions at each intermediate solution; for details see the *assessment delivery* section). The MP-QTI delivery system thus keeps track of which tasks an examinee has taken up through any point—always a contiguous block containing #11—and chooses the task to present next based on the value of the Observable Variable named Accuracy, that is, the final solution for the mixture. If it is satisfactory, the delivery system moves to the next harder task in the set of twenty-one,

and if it unsatisfactory, the delivery system moves to the next easier one, until a total of eleven tasks has been presented.

6.4 Materials and Presentation

The presentation of the assessment is one of the four main processes for *assessment delivery*. For the purposes of the demonstration, we developed presentation materials for delivering the assessment to examinees. The parts of the presentation include an introductory screen and the primary interaction screen (described in the *template* as Presentation and Materials), a crib sheet, presentation graphics (also described in the *template* as Presentation and Materials), and supporting databases (described in the *template* as Evaluation Data).

An introductory screen acquaints the examinee with the nature of the assessment. This screen names the powders and the reagents (tests). If the examinee is not familiar with the powders and reactions, the introductory screen will give descriptions and offer the examinee an opportunity to print these descriptions. For example, examinees may need to be reminded of which tests mask the results of others (e.g., because flour browns when heated and a mixture of flour and sugar will caramelize, heat does not necessarily indicate either the presence or absence of flour).

The primary interaction screen, shown in Figure 19, is the area in which examinees choose tests, receives test results, and indicate their deductions. The elements of this interaction screen are:

- An indication of the minimum and maximum numbers of elements in the Mystery Powder. These are set values in the *task specification* of the task being delivered.
- A pull-down list that includes the tests available and the potential combinations of elements in the Mystery Powder. The examinee uses the pull-down list to select the next test to perform or to indicate that he or she has arrived at a solution. This environment and these affordances are the same for all tasks.
- An area in which video or time-lapse sequences show the outcome of the test specified by the examinee.
- A narrative description of the outcome of the test.
- An optional crib sheet in which the examinee may make notes on tests and observations in order to reduce memory load.
- A set of buttons reflecting deductions the examinee has made with regard to the powders that are in the Mystery Powder, out of the Mystery Powder, or for which all six tests would not provide enough information to deduce whether or not they are present. The interaction of the examinee with this set of buttons produces the Work Products described above as a Solution Matrices—one after each test an examinee administered, and the final solution.

This interaction screen displays a visual outcome and may also display a message. For example, with the taste test of mixtures including sugar but not salt, the message of

"Tastes sweet" accompanies a picture of a young man eating cotton candy (see Figure 21). For the water test of mixtures including plaster, the message "The next day" accompanies a hand banging the solidified powder.

Figure 19. Prototype for Delivery of the Mystery Powders Task

Mystery Powders

You have been given a mystery powder consisting of at least one but not more than two powder components. You can determine the components given information from the available experiments.

Current experiment: Taste

Tastes sweet.

Step one: what can you deduce so far? Indicate whether each possible component is In or Out of mixture, or whether that cannot be known yet:

In	Out	Can't tell	Can't know	
o	o	o	o	Cornstarch
o	o	o	o	Flour
o	o	o	o	Plaster
o	o	o	o	Salt
o	o	o	o	Soda
o	o	o	o	Sugar

Step two: which experiment do you want to do next?

Heat Iodine Look Taste Vinegar Water

Do the indicated experiment

Done – My answer is shown above in Step 1

As the examinee engages with the main task, she is prompted to record her deductions following each test and to indicate either a final solution or her choice of a next test. In doing this, the examinee may wish to record notes. A crib sheet is available that serves as scratch paper; the examinee can scroll so that she may enter as much information as she wishes. After choosing a test, the examinee will be given observational results of the test in visual and verbal forms. The possible observations are listed in Figure 20. The examinee is prompted to note her deductions following each test—which powders she thinks are in the mixture, not in the mixture, cannot yet tell, and cannot be determined (even with all available evidence). If the examinee has not yet arrived at a final solution and wishes to collect more evidence, she can select another test to be run (Step 2 in Figure 19). When the examinee has reached her final solution, she indicates that she is done—that her deductions in Step 1 (see Figure 19) represent her final answer.

Figure 20. Possible Observations in Mystery Powders

Visual	Water	Iodine	Vinegar	Heat	Taste
No Obs.	No Obs.	No Obs.	No Obs.	No Obs.	No Obs.
Crystal	Dissolve	Not blue	No Fizz	Nothing	Tasteless
Powder	Gooey mess	Blue	Fizz	Brown	Sweet
Mixture	Lumpy/muddy			Caramelize	Salty
	Lumpy/hardens				Sweet and Salty

Given that this is a demonstration, we attempted to make it visually interesting. We used graphics and icons to supplement the text. The Mystery Powders assessment uses the following design features:

- There are moving graphics where applicable for chemical interactions. The effectiveness of 500K video clips depends on the speed of the Internet connection, compared with 30K for stills. Therefore, we leave it to the examinee to initiate videos by clicking on them.
- There are standard videos of reactions that take place in seconds or a few minutes, such as the fizzing of vinegar with baking soda.
- There are time–lapse moving images of slower reactions such as caramelization of sugar, browning of flour, etc.
- There is text to supplement all of the visuals.

We designed the Mystery Powders Simulation assessment to be available for reuse by the shareware community. This imposed a requirement that the components we use be shareware. The Internet, the XML markup language, and browser features were treated as givens. QTI (Question & Test Interoperability) specification for the XML exchange of question and test information was used to specify the format of messages, and the public domain MySQL database was used to implement the control of messaging.

To deliver MP-QTI, a large number of software components needed to interconnect. The assessment logic is deterministic; every possible examinee action can be evaluated according to previously computed results. A set of databases fully supports the logic of the delivery system:

- The *Observations* database indicates which observations will result from every combination of powders and tests.
- The *Deductions* database indicates what the examinee should be able to conclude from a given set of observations, minimum and maximum.
- The *Optimal Tests* database indicates, for each possible combination of prior observations, the relative power of each remaining test that the examinee might choose.
- The *Item Difficulty* database orders the 480 tasks (powder combination, minimum and maximum) by relative difficulty.

Furthermore, the four processes depicted in Figure 2 draw upon the aforementioned presentation materials such as photos, video clips, and tables. The materials and files are all specified in the *template* and *task specification*. In the implemented assessment, these materials reside in the *task/evidence composite library* shared by all processes, shown in the center of Figure 2.

7.0 Assessment Delivery via Four-Process Architecture

7.1 Overview

The last layer of assessment design is *assessment delivery*. The primary activities that occur in this layer are represented in the four-process delivery system - the phase during which the operational processes of the assessment are put into action. The four-process architecture for *assessment delivery* (Almond, Steinberg, & Mislevy, 2002) includes one process each for *activity selection*, *presentation*, *evidence identification* (or task-level scoring), and *evidence accumulation* (or test-level scoring). A central database library (the *task/evidence composite library*) serves to tie together information from each process. Examinees interact solely with the *presentation process* (see Figure 2).

For the MP-QTI demonstration, all four of the processes were implemented as web applications for easy access via Web browser. Each process can stand alone and even run on a separate computer, although the processes may communicate with a common central database that holds information about the assessment and the examinee. To promote interoperability of PADI assessments with other systems, we assembled presentation specifications and result scores into Question and Test Interoperability (QTI) XML documents as specified by IMS Global Learning Consortium, Inc. (2000). Thus, some of the communication implied occurs by transferring these QTI documents via the HTTP[10] protocol between Web applications.

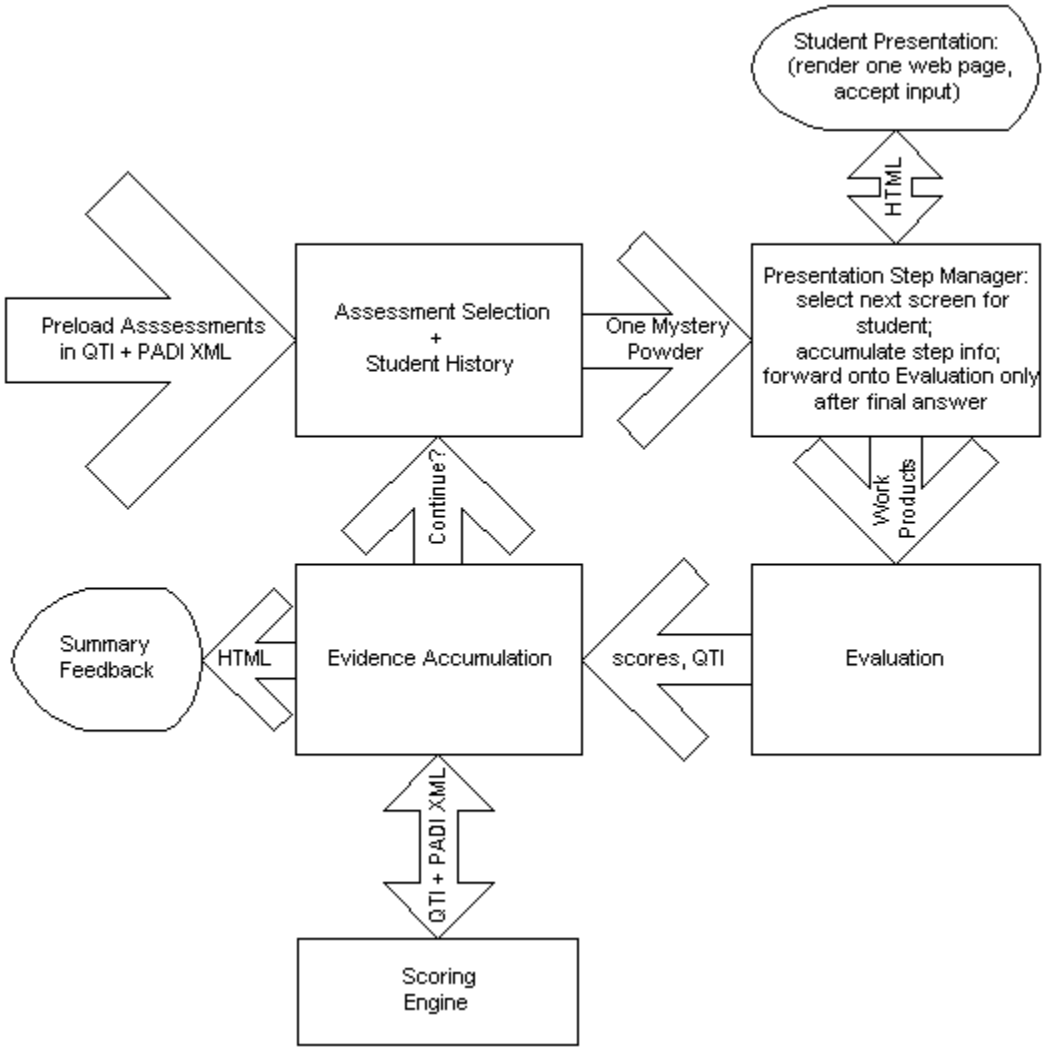
In a typical *assessment delivery* system, the four processes would be relatively invisible to the examinee—only the presentation of the task would be visible. However, to promote understanding of the different processes, this Mystery Powders implementation is designed to make visible a summary at the end of each process—an explicit statement of the transition from one process to the next. This summary Web page displays a graphic to identify the current process (with its oval colored red) and also shows the inputs received by the process and the outputs that are about to be sent to the next process. The outputs are displayed in HTML fields that allow editing so that a designer can manipulate the information flowing from one process to another to both test and understand the system.

The four-process architecture for the Mystery Powders Simulation is complex. Figure 21 depicts this architecture with additional constraints (compared with Figure 2). Prior to delivery, MP-QTI assessment tasks, as described in *task specifications* and authored in QTI and XML, are made available to the delivery system. The *activity selection process* initiates the assessment by selecting the initial task for presentation to the examinee. Because MP-QTI involves multi-step tasks (i.e., a sequence of tests of the mixture of powders), the *presentation process* must provide some kind of looping ability, allowing an examinee to run a number of virtual experiments before the process is signaled (via the examinee's final answer) that the presentation should finish. In Figure 21, the *presentation process* is called a Presentation Step Manager to emphasize that it has

[10] HTTP is an abbreviation for Hypertext Transfer Protocol, the standard protocol for the transfer of information between a browser and a web site, or in this case, between two web sites.

additional responsibilities and that it runs a local loop (confined to that rectangle in the diagram) until the examinee indicates that the final answer has been submitted.

Figure 21. Four-Process Architecture with Additional Constraints



Examinee Work Products are then passed onto the *evidence identification process*, which carries out the scoring algorithms detailed in the Evaluation Phase objects in the MP-QTI *template* (see Appendices A, B, and C for details). Task–level scores, in the form of OVs (Accuracy, Acuity, Efficiency, and Overall) are then passed on to the *evidence accumulation process*. In addition to saving the scores of the OVs, the *evidence accumulation process* composes and sends a QTI document to the Scoring Engine, then interprets the results in order to save posterior estimates of the examinee's abilities q_{DK} and q_{IS}. The bottom of Figure 21 amends the original diagram (Figure 2) by adding a communication channel between the *evidence accumulation process* and the BEAR Scoring Engine. This level of detail is hidden to the other processes and could be carried out by any program, or even by a human by hand, so long as it is able to receive the required input (OVs and task parameters) and produce the specified output (posterior distributions). In the MP-QTI demonstration, this communication with the Scoring Engine

is described in detail in the guide to the PADI Gradebook (Hamel, Mislevy & Kennedy, 2006). The *activity selection process* is engaged, and another task is selected via Lord's (1971, 1980) flexilevel adaptive testing algorithm as described previously until the examinee has taken a full assessment of eleven adaptively-selected tasks.

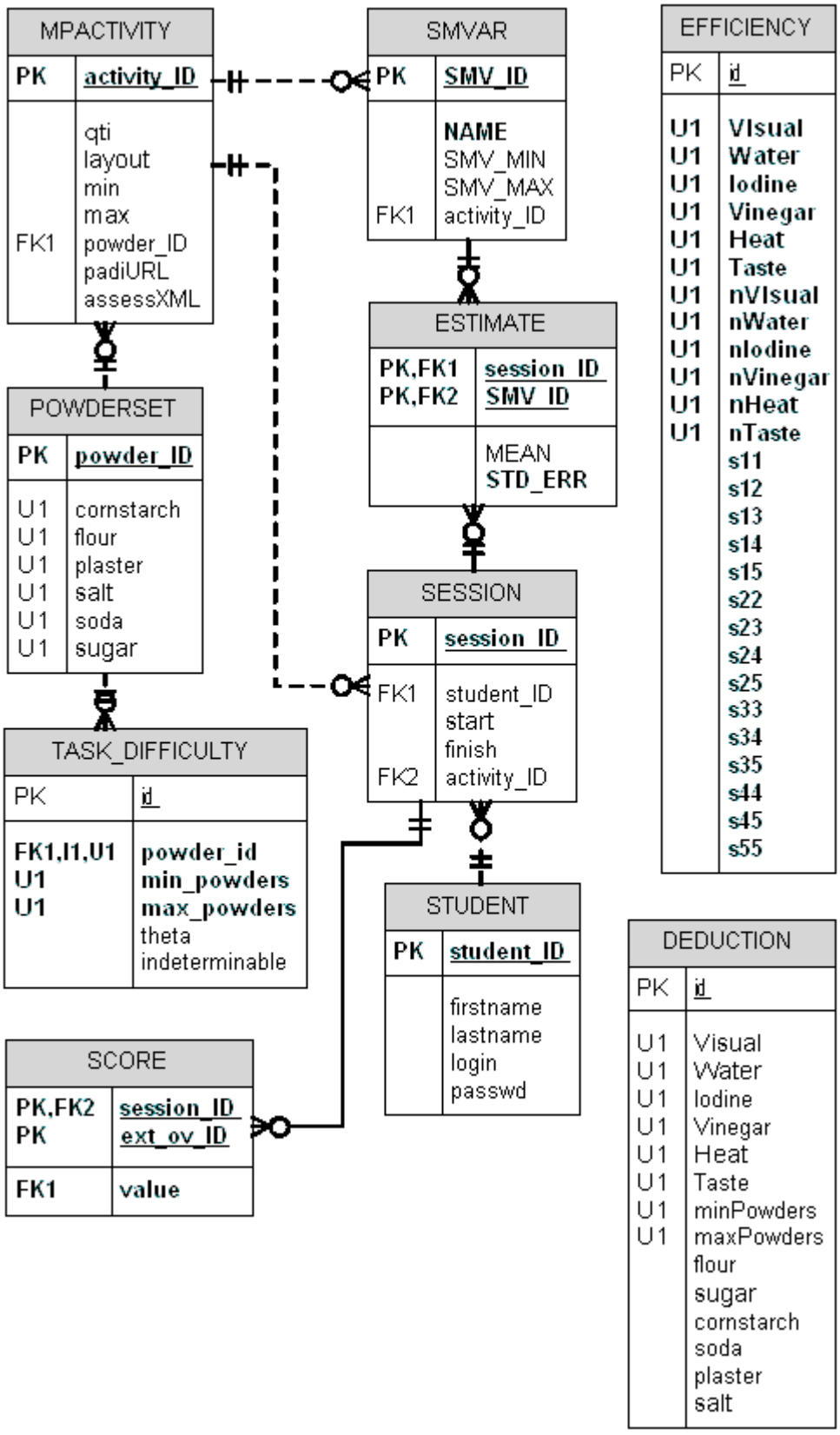
The central library of this four-process implementation is a database consisting of the entities and relationships shown in Figure 22. Descriptions of these database tables are provided in 5. In this figure, each table is represented as a rectangle, with the name of the table on a gray background, and the names of the fields of the table listed within the right-hand column of each rectangle. Fields that are primary keys are listed with an underline and also a "PK" notice in the left-hand column. Foreign-key relations[11] are indicated with arrows between tables. The foreign-keyed fields are annotated with an "FK" designation. Finally, fields that are indexed (besides PK and FK fields) are designated with a "U" for a unique index or an "I" for a simple index.

Table 5. Database Tables and Descriptions

Name of Table	Description
DEDUCTION	The inferences that an examinee is expected to make given experimental results; used to evaluate "Acuity"
EFFICIENCY	The quickest way, using the fewest experiments, to uncover the Mystery Powder, given a powder combination (e.g., start with the "Taste" experiment)
ESTIMATE	Examinee's proficiency estimate, provided by the Scoring Engine
MPACTIVITY	Assessment tasks, represented as an aggregation of powder, layout, and QTI information
POWDERSET	Combinations of powders
SCORE	The value of a given Observable Variable for a given examinee session
SESSION	Embodiment of an examinee's attempt to complete an assessment task
SMVAR	Storage for Student Model Variable definitions of the latent trait being assessed. Estimates for these variables are stored in the ESTIMATE table.
EXAMINEE	The identity of a given examinee

[11] A foreign key is a column in one table (the detail table) the contents of which match the primary key of a record in another table (the master table). For example, consider a master table of examinees, and a detail table of scores, where each score has a foreign key to the examinee's row in the examinee table.

Figure 22. Entity-Relationship Diagram of Four-Process Delivery System

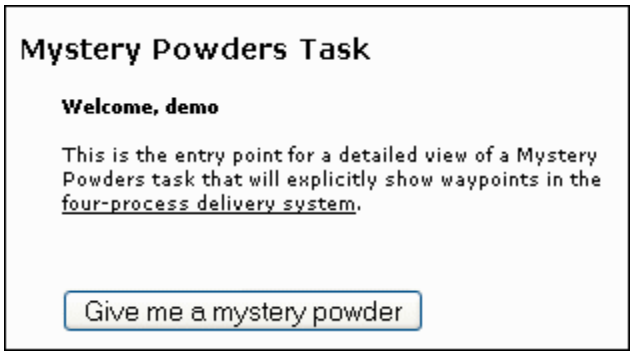


7.2 Activity Selection Process

The first process in the four-process architecture is the *activity selection process*. In general, any criterion could be used to decide which task to present to an examinee. In classroom practice, an experienced teacher would choose a task that is just challenging enough—not too hard and not too easy - for a particular examinee. In the MP-QTI implementation, the performance history of an examinee is taken into account in order to select an appropriate assessment task. Under Lord's flexilevel adaptive testing scheme, the middle difficulty task is chosen as the first task presented, and successive tasks are chosen to be easier after an incorrect solution and harder after a correct solution. The determination of success is a value of 6 or higher on the Bundled Observable Variable (see the *evidence identification* section below) that translates to either a final correct solution or mostly correct intermediate solutions. The *activity selection process* in MP-QTI is thus both automated and adaptive. It is automated because the algorithm chooses a task without human intervention; it is adaptive because the examinee will get a task that is more or less difficult depending on their previous performance.

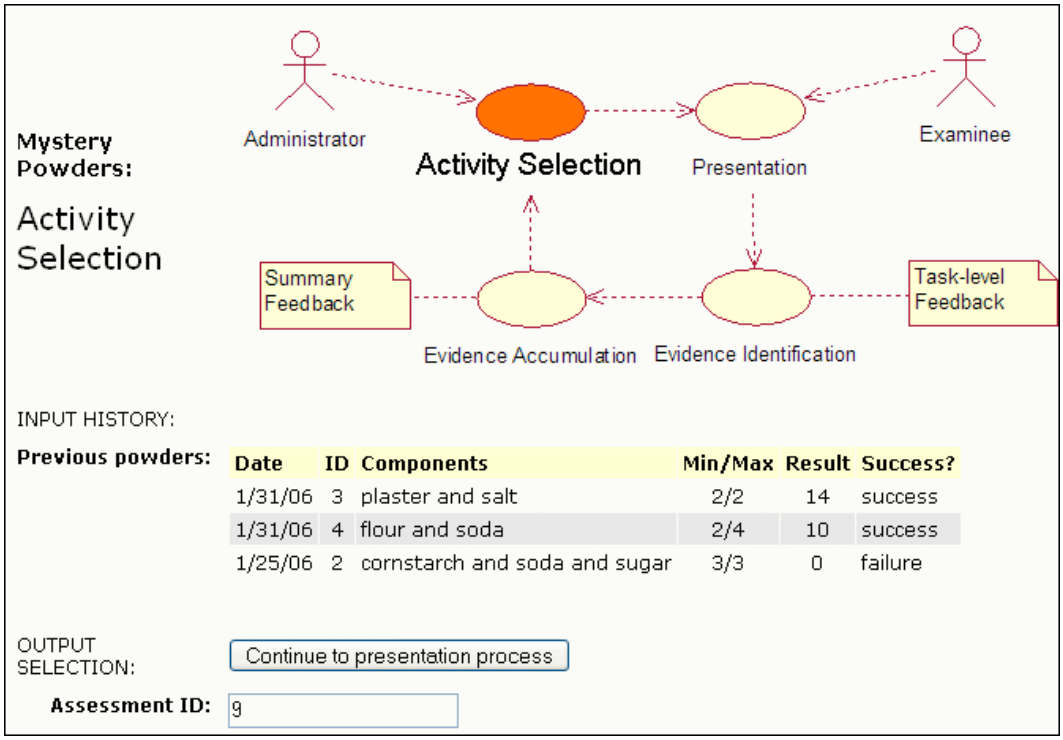
The welcome screen for MP-QTI is presented in Figure 23. This screen is shown to the examinee after their login. Clicking the button labeled "Give me a mystery powder" indicates that the examinee (named "demo") is requesting a new task. At that time, the summary of the examinee's performances on previous tasks is retrieved from the central library. This summary includes which tasks the examinee has taken already and what the scores were.

Figure 23. Welcome Screen for Mystery Powders Simulation



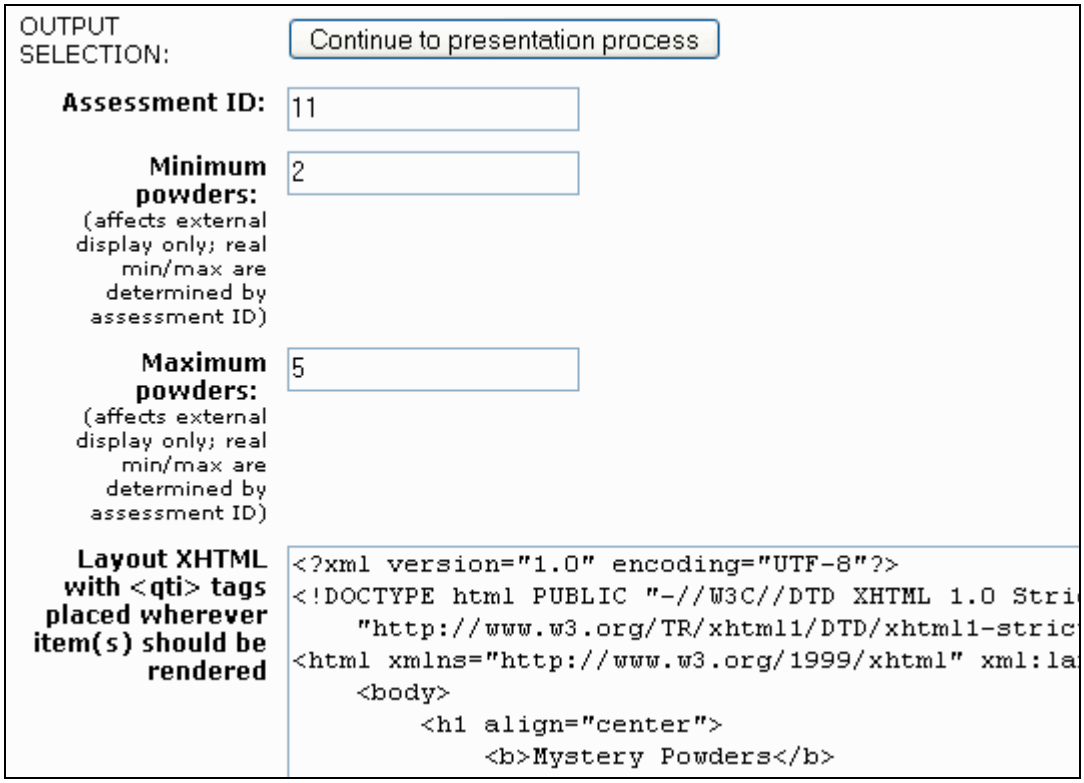
We developed summary pages to represent the culmination of each of the four processes. As mentioned, this would not be seen by the examinee in actual testing; it is provided to make explicit the inputs, outputs, and interactions of the processes in the delivery system. The input part of the summary page for the *activity selection process* is shown in Figure 24. It identifies the current process with a red oval and bold typeface within the original four-process diagram. Inputs to the selection process are the examinee's previous attempts, listed in the example in Figure 24 as three tasks (id numbers, powders, minimums, and maximums). The oldest performance listed on the figure (1/25/06) was a failure, and the more recent performances (both on 1/31/06) were successes. As a result, the next task selected will be slightly more difficult than most recent task (id number 3), as per the flexilevel algorithm.

Figure 24. Summary Page for the Activity Selection Process (Input Part)



The output part of the *activity selection process* summary page is shown in Figure 25. It identifies the selected task in terms of item id, minimum number of powders, and maximum number of powders. In Figure 25, some of the fields of the assessment task with ID number 11 are shown, including the minimum and maximum numbers of powders (2 and 5, respectively) and the layout XHTML. A designer can edit these values to change the task. Any change made should be internally consistent: a change in the assessment ID implies a change in the appropriate minimum, maximum, layout, etc. The layout and QTI fields are shown primarily to allow minor tweaks to details for testing or tuning processes; changing the ID implies a major change of powder components and experimental feedback.

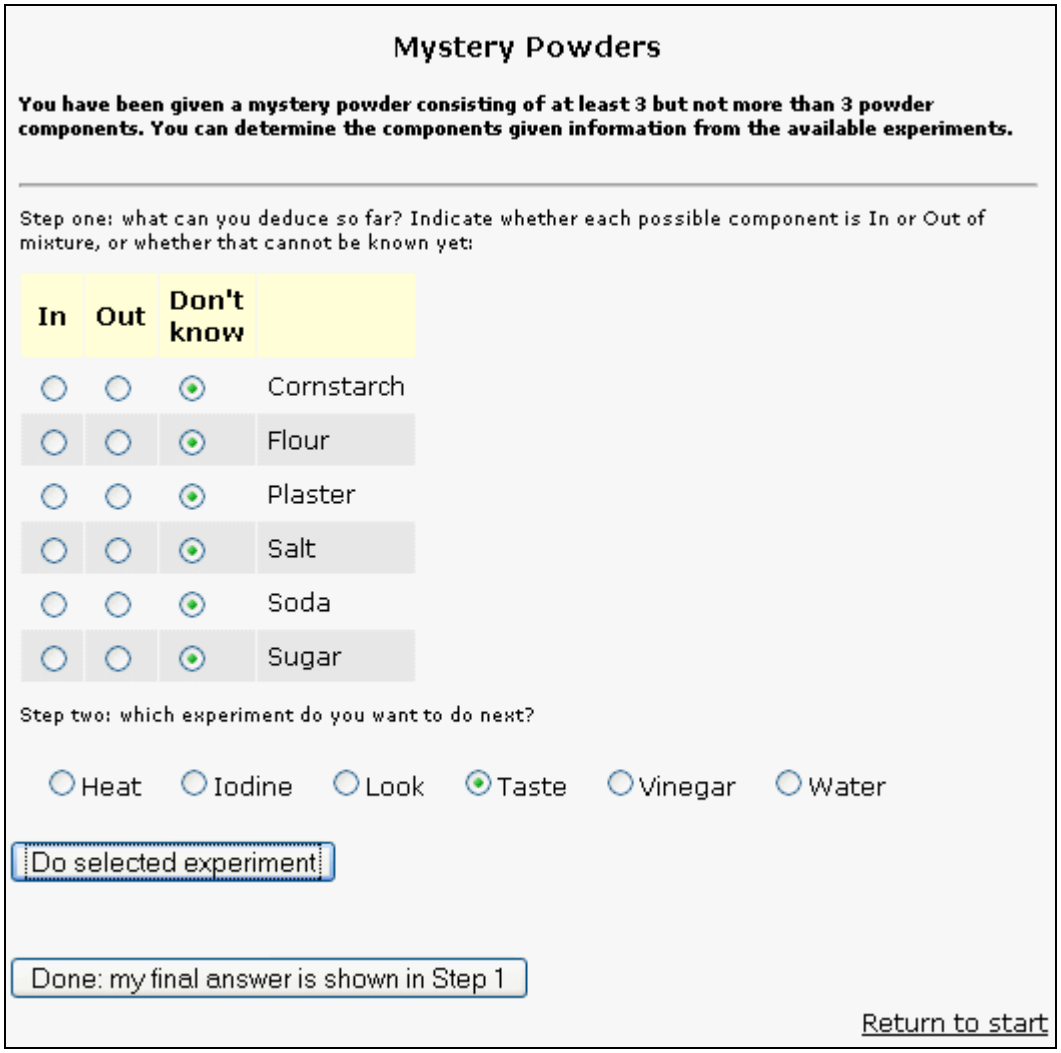
Figure 25. Summary Page for the Activity Selection Process (Output Part)



7.3 Presentation Process

The next process in the four-process architecture is the *presentation process*. This process for MP-QTI is fairly complex. As described in the overview, ascertaining the composition of a Mystery Powder typically requires that the examinee execute several virtual experiments before being able to draw a definitive conclusion. The series of screens presented to the examinee include an initial presentation screen (Figure 26) and screens including observational results from each test.

Figure 26. Initial Screen of Presentation



The initial presentation screen shown in Figure 26 is made up of several blocks. First, a statement of the problem at the top declares the minimum and maximum number of components in the powder mixture, along with an assurance that the mystery is determinable given the results of the available experiments[12]. The next part begins with a label for “Step One” and includes a solicitation of any conclusions the examinee can draw at the moment for each of the six potential powders. At this point, however, the examinee cannot draw any conclusions because no tests have been run; therefore, they would skip over this section on the initial screen. Below the Step One section is a Step Two section - a solicitation of the examinee’s selection of the experiment to perform. At the bottom of the initial presentation screen are two buttons that give the examinee the choice of either proceeding with the selected experiment or finishing the task by declaring that the

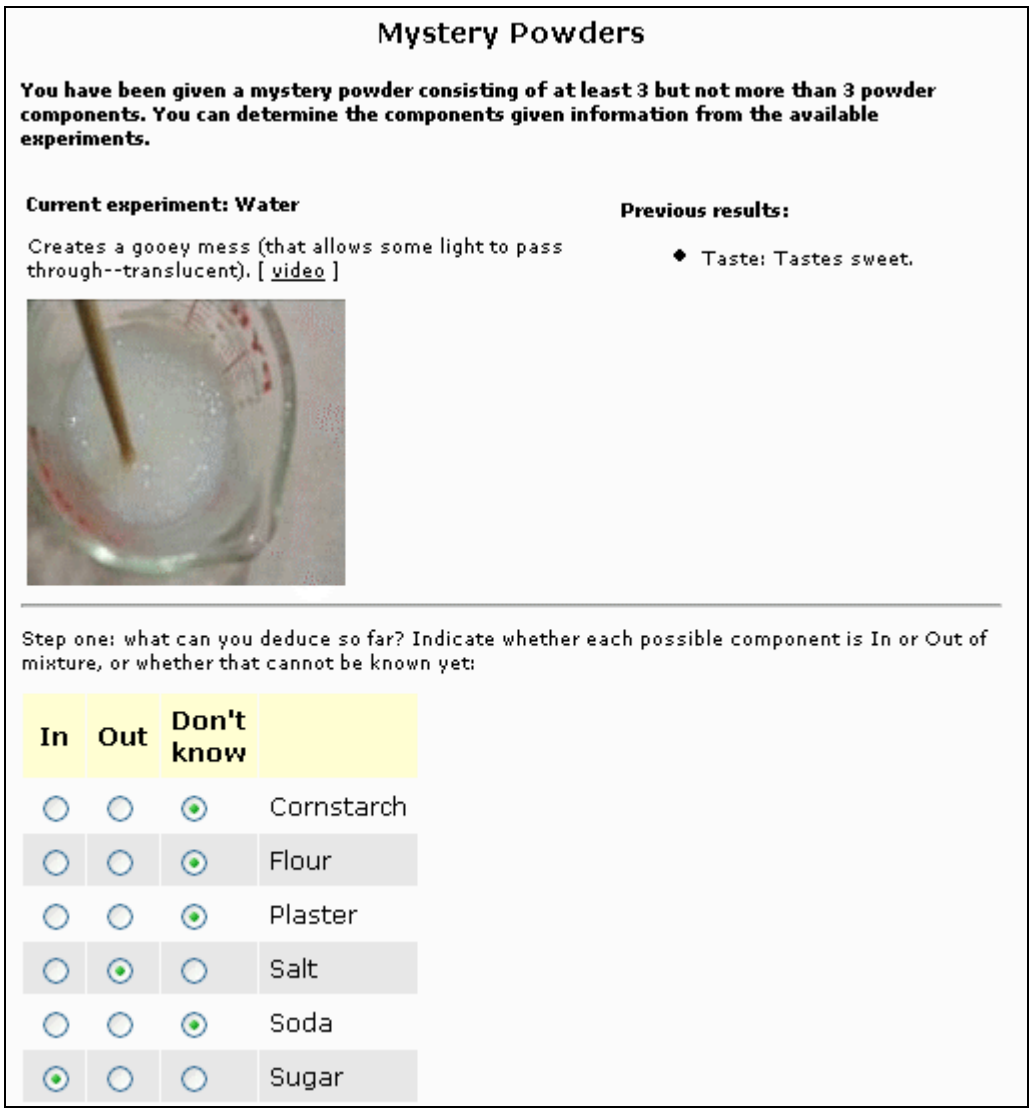
[12] Some powder combinations can be impossible to determine with the given experiments—they are indeterminable. In this MP-QTI implementation, we avoid giving indeterminable powders to examinees since such powders would add additional complexity to the already difficult tasks.

examinee has finished the task. Finally, in the bottom right–hand corner is a link to start over with a new powder mixture (new task).

Following the initial presentation screen, the examinee will receive a series of screens providing the results to each selected test, soliciting the examinee's deductions following each test, requesting additional test selections, and allowing for the identification of a final solution. Figure 27 provides an example of these presentation screens. In the top section of Figure 27, a pictorial representation of the results for the water experiment is shown. The water experiment, coupled with this particular powder combination, yielded a "gooey mess," as described in the text accompanying the picture. An additional hyperlink provides a video of the experiment, showing water being stirred into the powder. In the upper right part of the screen, the results of previous experiments are summarized; in this case, a previous taste experiment is summarized as "sweet" - indicating the presence of sugar and absence of salt. At the bottom of the screen (Step One) are six sets of radio buttons where the examinee can enter her deductions about the composition of the powder mixture.

The pictured display shows the examinee's conclusions from the previous taste experiment, but none yet from the water experiment. Using the available evidence from the water experiment, the examinee could deduce that the "gooey mess" result implies cornstarch is in the powder combination and that plaster and flour are not in because these would mask the gooey result by resulting in a lumpy/muddy result and a lumpy/hard result, respectively. This recording of the examinee's deductions (which comprise a Solution Matrix Work Product, as described in the *template*) is followed by another screen offering Step Two—the selection of an additional test to perform or choice to specify their deductions as final results.

Figure 27. Presentation Screen (Continued)

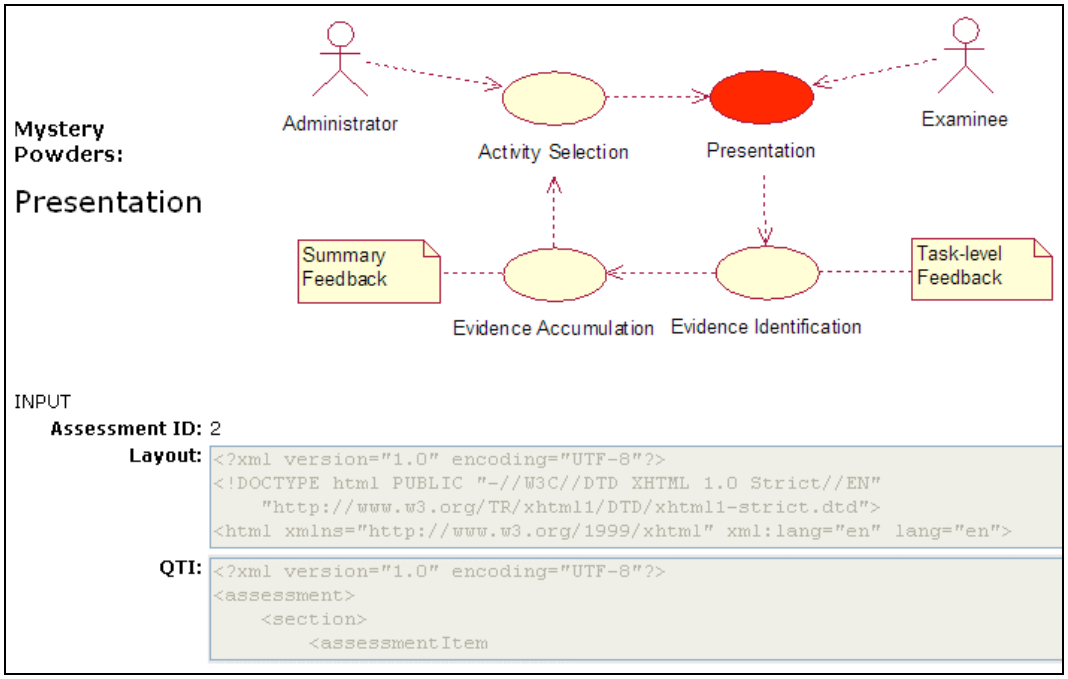


In addition to use of experiment evidence, the statement of the minimum and maximum number of components (at the top of Figure 27) may help narrow the possibilities for the examinee. In the case depicted in Figure 27, the deductions that cornstarch is in and plaster and flour are out (from the water experiment) and that sugar is in and salt is out (from the taste experiment) imply that only one potential component, soda, remains undecided. Because two components can be known to be in, three components can be known to be out, and the minimum number of components is specified as three, it can be deduced with no additional tests that soda is in the mixture. This is an illustration of how the selected minimum and maximums for the given powder combination impact the difficulty of tasks. Given the same powder combination, tasks are harder when the difference between the minimum and maximum is greater.

The summary of the *presentation process* in Figure 28 shows all the inputs that come into the presentation, including the layout and QTI XML that was provided by the *activity selection process*. These inputs are shown on the summary page for information only and

cannot be changed by a designer because they already have been used during the selection process.

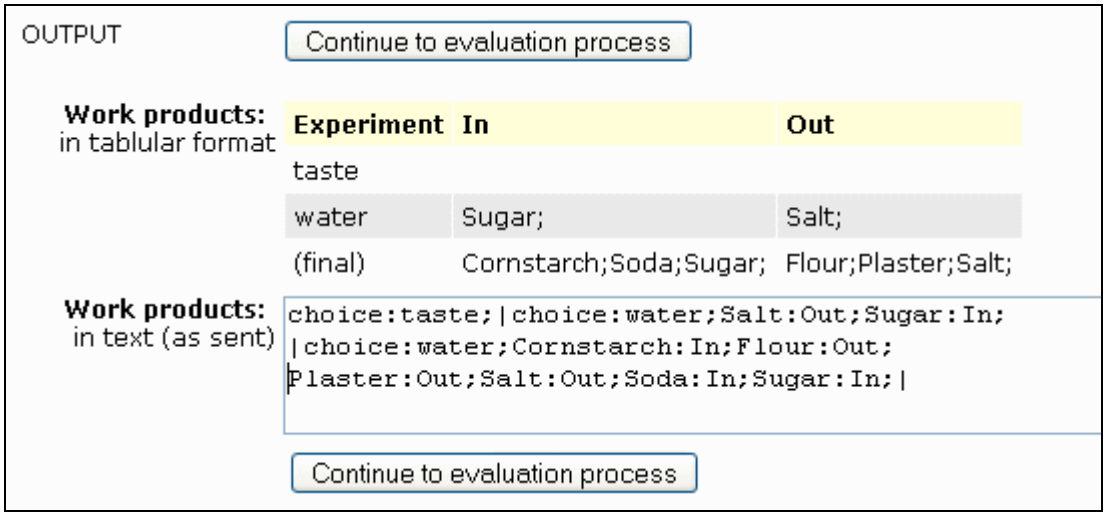
Figure 28. Summary Page for the Presentation Process (Input Part)



The *presentation process* creates an output to supply to the next process, *evidence identification* or task-level scoring. The output is a summary of all the choices made by the examinee, or the examinee's Work Products, in the form of the vector of tests selected and the Solution Matrix. Figure 29 shows a summary page for the output part of the *presentation process*. Two views of the same output (Work Product) are shown; the upper part of this output summary is in tabular format and the lower part is a string of text in an editable form. The editable form constitutes the actual document transmitted to the *evidence identification process*. In this example, the tabular form is most readily interpretable: in the first line, the examinee chose to see the results of the taste experiment. When the examinee selected this initial experiment, no potential powders were declared in or out of the Mystery Powder mixture. In other words, before the results of the first experiment are obtained, there are no conclusions possible (without guessing).

The next line in the table, shown with a gray background in Figure 29, represents the examinee's selection of the water experiment as well as the examinee's deductions following the taste experiment—that sugar is in and salt is out. The final line indicates that the examinee made a final set of deductions (cornstarch, soda, and sugar are in; flour, plaster, and salt are out) by clicking the button (in Step Two) to indicate a final answer. This summary screen provides a means for an assessment designer to manipulate the answers (Work Products) in text form. Only the text shown in the editable box is transferred to the *evidence identification process*.

Figure 29. Summary Page for the Presentation Process (Output Part)



7.4 Evidence Identification Process

The third process in the four-process architecture is *evidence identification*. The previous section on the Evidence Model (part of the CAF layer) offered some information pertinent to this process; here we go into more detail. The *evidence identification process* for MP-QTI uses the examinee's Work Products as input to a scoring process. This results in three intermediate Observable Variables that are then combined into a final, bundled OV. The scoring requires relatively complex algorithms because there are many potential paths through the six available tests. The evaluation algorithm was developed to measure both the Efficiency of the chosen path as well as the Acuity of the conclusions made by the examinee during the path traversal. For example, if an examinee first chooses the water test and gets a result that identifies cornstarch in the powder, it would be redundant and inefficient to ask for the iodine experiment. The result would be a foregone conclusion: the result would turn blue because the mixture contains cornstarch. Therefore, the Efficiency score would be low. Likewise, if the water result makes it possible to identify cornstarch in the powder, the evaluation algorithm will only result in the highest Acuity score if cornstarch is identified by the examinee.

Three intermediate OVs, namely Acuity, Accuracy, and Efficiency, result from the evaluation of Work Products. Acuity is a measure of the completeness and accuracy of the deductions that an examinee makes after each experiment, except for the final answer (the Accuracy OV is based on the final answer.) For example, if the examinee asks for four experiments, Acuity will measure the quality of the deductions after the first three experiments. The Acuity algorithm (see Appendix B) takes each experiment in turn, determining how many of the potential components have been correctly identified after each test, and assigns a percentage of correctness, dividing the number correctly identified by the total number powders. The Acuity algorithm then averages all of these percentages for a final percentage and summarizes them on a four-point scale with increasing percentage ranges.

Accuracy is a dichotomous OV representing the correctness of the final answer, with score of 1 for correct and 0 for incorrect (see Appendix A). Efficiency is a measure of how efficiently an examinee picked each experiment. The Efficiency algorithm (see Appendix C) has a database of all possible paths or all possible selections of experiments after any given set of experiments. Each possible path has been ranked for Efficiency, a priori. The stated minimum and maximum play an important role in deciding the most efficient next experiment. Efficiency is also a dichotomous OV – a score of 1 is given for perfect Efficiency and a score is 0 is given for less than perfect Efficiency.

After the intermediate OVs (Acuity, Accuracy, Efficiency) have been scored by their respective algorithms, the three scores are combined into a final OV called the bundled Observable Variable. We create this final score to take into account the dependencies among the three intermediate OVs. This combination is done using the mappings in Table 6. Arithmetically speaking, the bundling formula is: bundled score = accuracy * 8 + acuity * 2 + efficiency. These are not weights in any psychometric sense; they are simply indices in the coding scheme that maps triples of values from three variables into values of a single variable.

Table 6. Mappings for the Final Bundled Observable Variable

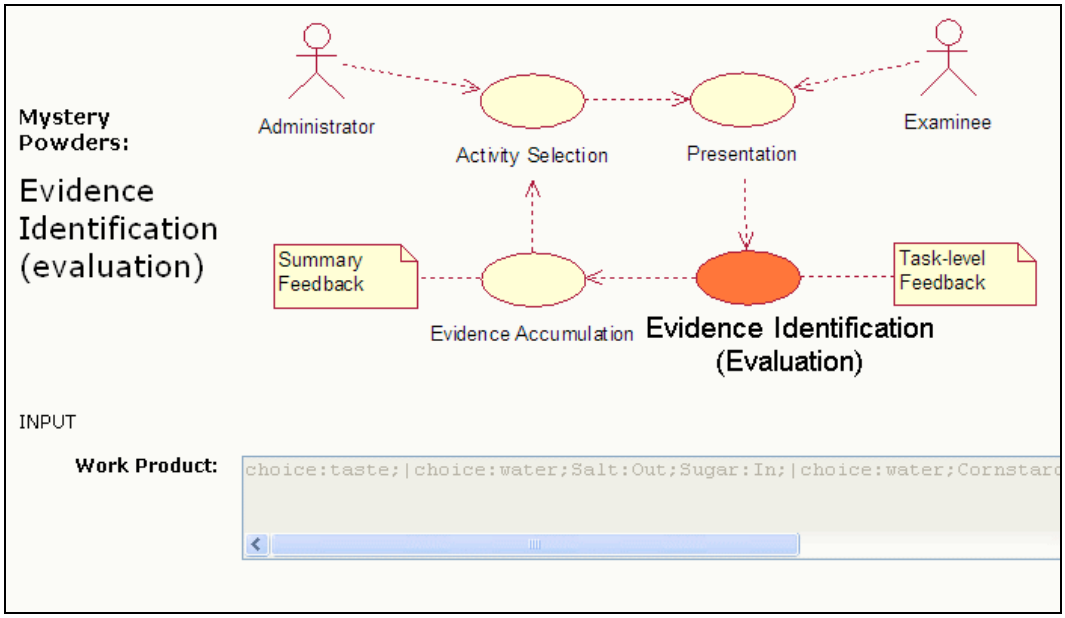
Bundled score	Accuracy	Acuity	Efficiency
0	0	0	0
1	0	0	1
2	0	1	0
3	0	1	1
4	0	2	0
5	0	2	1
6	0	3	0
7	0	3	1
8	1	0	0
9	1	0	1
10	1	1	0
11	1	1	1
12	1	2	0
13	1	2	1
14	1	3	0
15	1	3	1

The bundled OV is an output that is provided to the *evidence accumulation process*. The bundled OV is also incorporated into a QTI Results document supplied to the *evidence accumulation process*. An additional final OV is created, a true/false designation of whether the examinee's bundled score was above a certain threshold. As mentioned in the discussion on adaptive testing, this true/false designation serves as a declaration of whether the examinee has succeeded or not, as used to determine whether they should

receive a harder or easier subsequent task. The cutoff value for success is 6. Examinees with a score of 6 or more on the bundled OV are judged as having succeeded on a current task, and will be given a more difficult next task; examinees with score of 5 or less are judged as having failed on a current task and will receive a an easier next task.

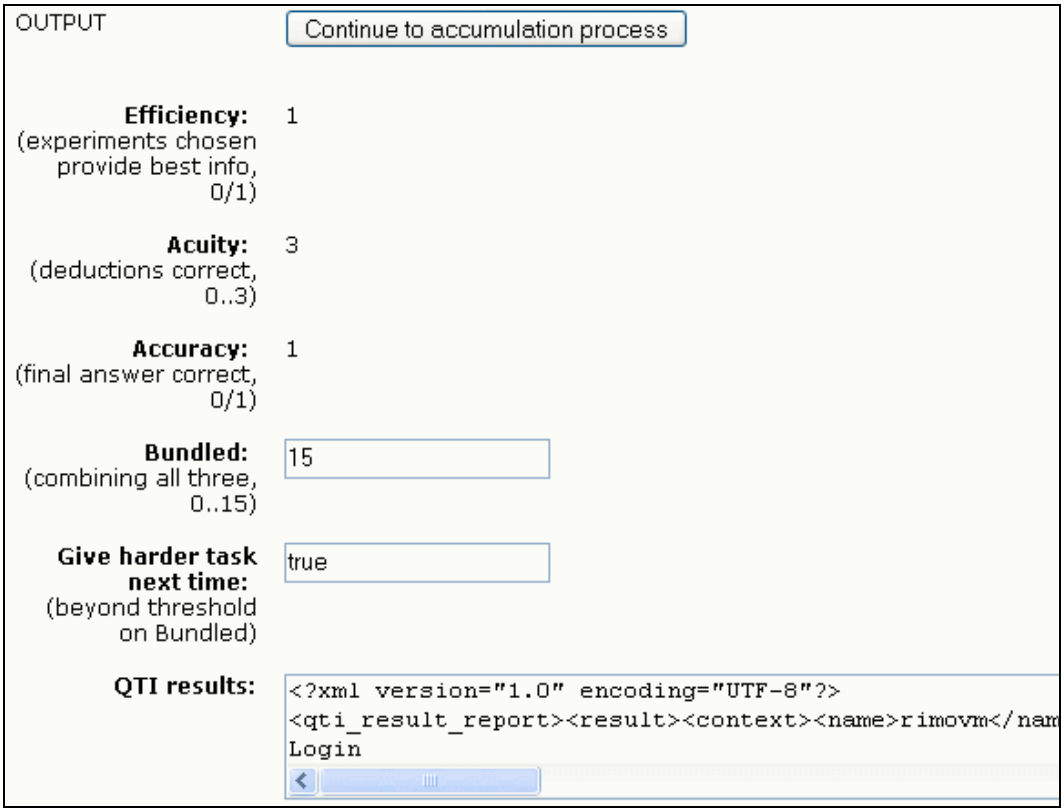
The summary of the *evidence identification process*, in Figure 30, shows the single input that came into this process, namely the string that summarizes the entire Work Product (selections of experiments and deductions following each test). This input, shown on the summary page, is for information only and cannot be changed because it already has been used during the process just completed.

Figure 30. Summary Page for the Evidence Identification Process (Input Part)



The output part of the *evidence identification process* in Figure 31 shows several pieces of information, including the values of the intermediate OVs, the value of the final OV, the determination of harder or easier subsequent tasks, and QTI results. The values for the intermediate OV indicate that the examinee received full credit for each of the OVs: a '1' for Efficiency (each test chosen was the single most efficient), a '3' for Acuity (examinee deductions were 100% correct) and a '1' for Accuracy (the final answer was completely accurate). These values are listed as read-only since they are not passed on to the accumulation process. Following this, the value of the bundled OV is listed as '15' = 8 * Accuracy + 2 * Acuity + Efficiency. Since the bundled OV score is 6 or above, a harder task is indicated for the next task. Finally, some QTI code is presented.

Figure 31. Summary Page for the Evidence Identification Process (Output Part)



In Figure 31, the three outputs will be transmitted to the accumulation process: the bundled OV, the pass/fail determination (whether a harder task is given the next time), and the QTI results document. These are all provided as editable fields. Although their information is somewhat redundant, we provided them separately so that the thresholds and algorithms are centralized in the *evidence identification process*. The *evidence accumulation process* will store these values and communicate with the Scoring Engine.

7.5 Evidence Accumulation Process

The final process in the four-process architecture is the *evidence accumulation process*. The *evidence accumulation process* for MP-QTI accomplishes a number of functions: it saves the examinee's final score (the bundled OV) and communicates with the Scoring Engine by sending it a QTI document and receiving back a posterior update of estimated examinee proficiency.

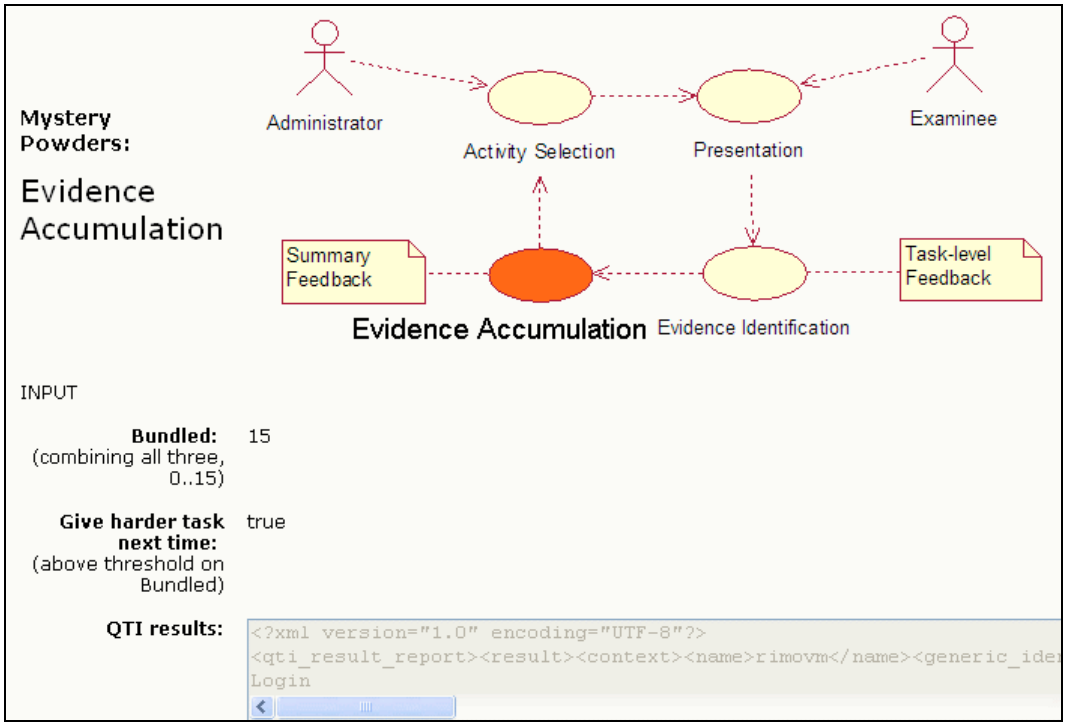
The *evidence accumulation process* combines a QTI document (received from the *evidence identification process*) with the XML description of the assessment task found in the central library. As described in detail in Hamel, Mislevy, & Kennedy (2006), the assessment XML document includes information about how OVs are expected to contribute information to SMVs, along with the calibration parameters needed for these calculations. To generate the QTI document for the assessment, the PADI design system *template* prompted the designer for various types of information. In the case of the Mystery Powders Simulation, the scoring matrix (shown earlier as in Figure 29) was

stored in the Measurement Model for the bundled OV. Figure 29 reflects designers' decisions that the domain knowledge SMV influences the examinee's performance on the Accuracy and Acuity OVs, while the inquiry skill SMV influences performance on the Efficiency and Acuity OVs. From Table , we can see that a final bundled score of 11 reflects 1 for Accuracy, 1 for Acuity, and 1 for Efficiency.

The calibration parameters and design matrix are also entered into the PADI design system (see the *template* in Figure 7). The PADI design system generates an XML document as expected by the Scoring Engine. The Scoring Engine uses this document and the QTI scoring document to update estimates about the examinee. The likelihood function induced by these values, through the MRCML model and the task's difficulty parameters, are combined with the prior distribution for q_{DK} and q_{IS} to produce an updated posterior distribution for the examinee's abilities. After receiving these results, the *evidence accumulation process* stores the updated estimates.

The summary of the *evidence accumulation process* in Figure 32 shows the inputs from the *evidence identification process*. These inputs are a final bundled OV score (15, in this example), determination of whether to administer a next, harder task (based on a threshold of '6' on the bundled OV), and some code for QTI results. These inputs are shown on the summary page for information only and cannot be changed because they already have been used during the *evidence identification process*.

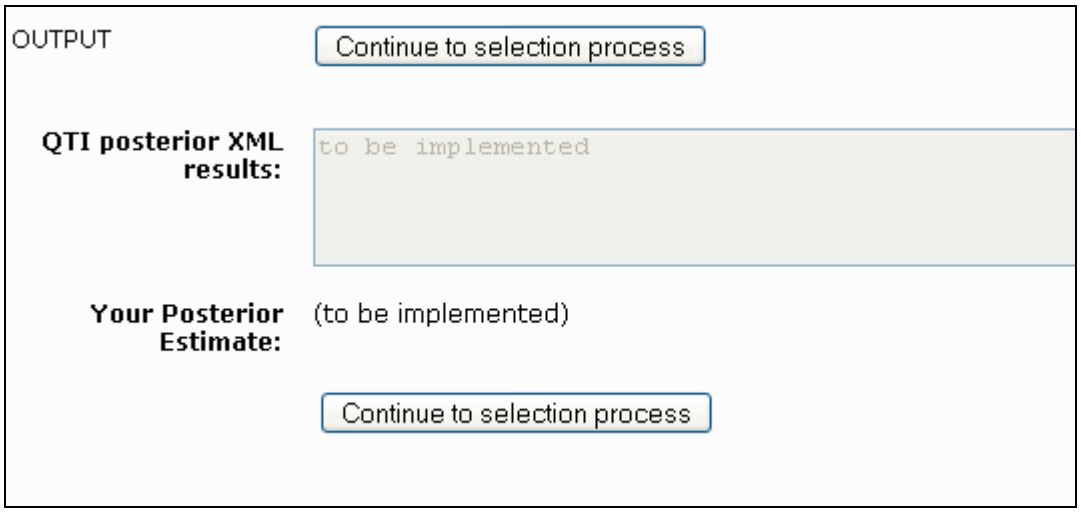
Figure 32. Summary Page for the Evidence Accumulation Process (Input Part)



Many of the outputs of the *evidence accumulation process*, shown in Figure 33, are not seen and involve saving values in the database. The exceptions are the two outputs listed (that are a result of communication with the Scoring Engine): the QTI posterior XML results and 'Your Posterior Estimate.' The Scoring Engine provides an update of the

posterior (most up-to-date) estimate of examinee abilities. This estimate is also saved in the database and is not passed on to the *activity selection process*. It is listed as an output because it is a result of the *evidence accumulation process* although it is not actually passed on to any other process.

Figure 33. Summary Page for the Evidence Accumulation Process (Output Part)



Because the *activity selection process* was designed to be able to act at the very beginning of an examinee session, it does not require information communicated via HTTP from another process. Therefore, unlike other processes, the *evidence accumulation process* does not forward outputs via HTTP. Instead, the *evidence accumulation process* saves the output information to the central database, and the subsequent *activity selection process* gathers all of its information from the central database.

8.0 Conclusion

The Mystery Powders–QTI demonstration project illustrates the use of the PADI design system object model, both for expressing the blueprint for assessment tasks and for implementing an assessment in a four-process architecture for its delivery system. Among the features it exhibits are computer-based simulation tasks, Web delivery, IMS/QTI compatible expression of assessment objects, and a multivariate Measurement Model with multiple, conditionally–dependent observations. As such, MP–QTI serves as an example for using the PADI design system with complex computer-based assessments.

The work illustrates distinctions among kinds of work that take place at different layers in the assessment enterprise, and shows how PADI design objects support the work within and across layers.

Domain analysis is where studies of research, existing practice, studies in the substance area, teacher experience, and so on contribute information that will be important in designing an assessment. In the Mystery Powders project, *domain analysis* focused on finding and studying a number of versions of Mystery Powders tasks that have been used in various instructional and assessment settings. We then reverse engineered these tasks to better understand design decisions that the originators of these examples had made and to inform decisions we would make for our demonstration.

The *domain modeling* layer organizes information gleaned from *domain analysis* into structures that reflect assessment arguments. In particular, we built two PADI *design patterns* that highlighted in narrative terms the key elements of an assessment argument, namely, the Knowledge, Skills and Abilities that are of interest, the kinds of examinee behaviors or performances one might observe to provide evidence, features of task settings that are necessary to make it possible to get the evidence, and features of tasks that can be varied to make tasks easier or harder, shift emphasis on what knowledge is emphasized, require different equipment or circumstances, and so on. One *design pattern* was quite general and addressed hypothetico-deductive problem-solving in finite problem spaces. This design pattern applies not only to Mystery Powders but also to a wide range of tasks that could be created for other domains and other educational levels, from simple children's games like Twenty Questions to troubleshooting hydraulics systems in jet airplanes. This *design pattern* supports the creation of assessment tasks that reflect inquiry standards as presented in authoritative guides such as the National Science Education Standards (National Research Council, 1996)—standards that may be applied across a wide range of domains in science and across grade levels.

The second *design pattern* focused on Mystery Powders tasks and could be used to help assessment designers create Mystery Powders tasks of various types, including lab experiments, multiple-choice versions, or the computer-based simulations that were built for this demonstration. This *design pattern* points out both what is common to all such tasks—each an exercise in hypothetico-deductive reasoning and, in particular, in the setting of analyzing mixtures of powders—and what may be varied within these parameters to accommodate local requirements of resources, constraints, and assessment purposes.

The *conceptual assessment framework* (CAF) lays out more technical blueprints for tasks or families of similar tasks. PADI provides structures called *task templates* and *task specifications* for this purpose. *Task templates* are schemas with slots for the elements specified in the Mislevy, Steinberg, and Almond (2003) *conceptual assessment framework*, namely Student, Evidence, and Task Models. Filling in these slots with additional structures describing activities, psychometric models, stimulus and work product descriptions, and evaluation rules creates blueprints for authoring many tasks with the same evidentiary structure and assessment argument. The Mystery Powders example illustrates the kind of information that appears in a *template*, and the thinking that carries the designer from a narrative argument suggested by a *design pattern* to the detail structures of actual tasks. *Task specifications* fill in the slots of *templates* to specialize the blueprint to individual tasks, in effect becoming a specification for authoring the implied task.

Of particular interest in the CAF is the specification of activities that suit interactive investigation in a computer–based simulation environment, evaluation rules that enable automated scoring in this environment, and a complex psychometric model—a multivariate item response model with multiple categorical responses and conditional dependence of observable variables within tasks.

The *assessment implementation* layer is where the elements of tasks specified in *task templates* are authored, details of evaluation rules are implemented, presentation materials are assembled, and psychometric parameters are put in place. The Mystery Powders example illustrates how all of these jobs are guided by the structures and the contents of the design objects created in the CAF. Because of how the objects in the CAF were created, the elements of tasks accord with the assessment argument laid out from the beginning of the process. In particular, the calibration of the multivariate psychometric model mentioned above was carried out with simulated data using the BEAR Scoring Engine developed in the PADI project.

The *assessment delivery* layer concerns the actual interaction of students with tasks. The Mystery Powders demonstration illustrates *assessment delivery* in accordance with the four-process architecture for *assessment delivery* systems (Almond, Steinberg, & Mislevy, 2002). The Mystery Powders delivery system is a fully functional Web-based system that optionally makes visible the messages, inputs, and outputs that the processes are sending one another. It is shown that the structure and nature of content of the HTML-based messages was specified in the *task template* and *task specifications*. Further, both the delivery system model and the form of the messages are compatible with the IMS/QTI international standards for the interoperability of digital assessment objects and services.

The Mystery Powders *presentation process* presents materials to students, in this case in an interactive simulation on a computer in an investigation that can take up to six experimental tests of a Mystery Powder. It manages whatever interactions are required of the student and captures their Work Products in the forms specified in *task templates* and *specifications*. In this case, several Work Products are captured—namely the sequence

of experimental tests a student carries out and her judgments after each test as to what powders may be ruled in or ruled out.

Next, the *evidence identification* or task-level scoring process in Mystery Powders illustrates an automated scoring procedure that operates on the multiple Work Products from each task noted above, and produces a vector of Observed Variables to pass to the next process, evidence *accumulation*. These response data are packaged and sent to the *evidence accumulation process* using the PADI-developed supporting program called Gradebook.

The *evidence accumulation* or test-level scoring process synthesizes across tasks the information about Student Model Variables that is contained in Observables Variables. The BEAR Scoring Engine, also developed in the PADI project, is used to update the multivariate Student Model using Bayesian procedures with the item parameters estimated during implementation.

The *activity selection process* in Mystery Powders also takes advantage of computer-based presentation by using an adaptive testing algorithm. The particular algorithm used in this demonstration is Lord's (1971) flexilevel testing procedure[13], in which a student moves to the next harder or next easier task in a predetermined sequence according to whether the previous task solution was successful or unsuccessful.

In sum, the Mystery Powders–QTI demonstration serves several purposes for several audiences. It is a nontrivial teaching example of the principles and practices of evidence-centered design for students and practitioners. It provides a meaningful example of the use of the PADI object model and design system that may be of interest to assessment designers. It shows the deep interconnections among narrative, substantively-grounded assessment arguments, specifications for the technical details of the operational elements of assessments, and the processes and operations of *assessment delivery*, all in terms of the structures of the PADI framework. This use is of particular interest to assessment designers and measurement specialists. In conjunction with the operational assessment system, it is an exemplar for developing other assessment systems based on the four-process architecture, in particular with specifications in terms of the PADI framework and with messaging consistent with IMS/QTI standards. System designers and database managers will find this aspect of the demonstration instructive. At this writing, Mystery Powders–QTI stands as the most complete exemplar to date of the use of the PADI in terms of design framework, implementation, and operation in a fully computerized assessment system.

[13] Our use of the flexilevel procedure for Activity Selection is an homage to Frederic Lord, in appreciation of his incomparable contributions to educational measurement.

References

Adams, R., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

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

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

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

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Hamel, L., Mislevy, R. J., & Kennedy, C. A. (2006). *A guide to the PADI gradebook* (PADI Technical Report 12). Menlo Park, CA: SRI International.

Hornke, L. F., & Habon, M. W. (1986). Rule-based bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369-380.

House of Representatives, Committee on Science and Technology. (October 29, 1986). *Investigation of the Challenger Accident, House Report 99-1016*.

IEEE-USA Board of Directors (2003). *Reverse engineering.* Position paper of the Institute for Electrical and Electronics Engineers, Inc., Washington, D.C

IMS Global Learning Consortium, Inc. (2000). *IMS Question & Test Interoperability specification: A review* (White Paper IMSWP-1 Version A). Burlington, MA: Author. Retrieved May 1, 2004, from http://www.imsglobal.org/question/whitepaper.pdf

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin,112*, 527-535.

Lord, F. M. (1971). The self-scoring flexi-level test. *Journal of Educational Measurement, 8,* 147-151.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed.) (pp. 13-103). New York: American Council on Education/Macmillan.

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

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J. (2003). Argument substance and argument structure. *Law, Probability, & Risk*, *2*, 237-258.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., Haertel, G., et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.

Mislevy, R. J. & Riconscente, M. M. (2005). *Evidence-centered assessment design* (PADI Technical Report 9). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-67.

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

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

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Rumbaugh, J., Jacobson, I., & Booch, G. (2004) *The Unified Modeling Language reference manual* (2nd ed.). Reading, MA: Addison Wesley.

Stanford Education Assessment Laboratory (2004). *Assessments/instruments, Mystery Powders-end of unit assessment, scoring instructions*. Retrieved September 20, 2006 at http://www.stanford.edu/dept/SUSE/SEAL/

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum.

Shute, V. J. & Torreano, L., & Willis, R. (2000). DNA: Towards an automated knowledge elicitation and organization tool. In S. P. Lajoie (Ed.) *Computers as Cognitive Tools, Volume 2.* Hillsdale, NJ: Lawrence Erlbaum, pp. 309-335.

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.

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

Wainer, H. (2000). *Computerized adaptive testing: A primer*. London: Erlbaum.

Wigmore, J. H. (1937). *The science of judicial proof* (3rd edition). Boston: Little, Brown, & Co.

Wilson, M. & Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181-198.

Wu, M., Adams, R. J., & Wilson, M. R. (1998). *ConQuest* [computer program]. Camberwell, Victoria: Australian Council for Educational Research.

APPENDIX A

Appendix A: Evaluation Procedures to Determine the Accuracy Observable Variable

An examinee's final Accuracy score is determined as follows:

1. The examinee's final Work Product serves as input to the evaluation algorithm
2. For the given task, the Work Product string is compared to the "optimal" deductions for a given task, using an evaluation algorithm.
3. An output OV is created with 1 = right and 0 = wrong

(1) Work Product Input to the Evaluation Algorithm

For any given Mystery Powders task, following each experiment of the mixture, the examinee decides whether each of the 6 potential powders is in, out, or unknown, yielding a set of 6 decision responses. When the examinee indicates their response is final, this set of decisions is evaluated as their final response. For each powder, the system codes decisions as: in = 1, out = 0, and don't know = N. The system codes responses in the order of flour, sugar, cornstarch, soda, plaster, and salt—resulting in the 6–digit response string. Thus, a final response string of 'N01110' indicates that the examinee did not know if flour was in the mix, decided that sugar and salt were not in the mix, and decided that cornstarch, soda, and plaster were in the mix.

(2) Evaluation Algorithm: Comparison of Examinee and Optimal Decisions

Accuracy is the quality of deductions for an examinee's final solution to a Mystery Powders task. Determination of Accuracy involves a table within a spreadsheet named 'deductions_2FIX.xls' (linked to in the PADI object model as Evaluation Action Data). The SQL table within the spreadsheet provides all possible accurate deductions for every set of experiments and results given all possible minimums and maximums. There are 6 possible experiments that can be run and 2^6-1 = 63 different combinations of experiments (excluding the trivial case of running no experiments at all). Table A.1 indicates the possible observations for each of the 6 experiments. There are 1,062 combinations of observations, given that some combinations cannot occur (e.g., a mixture that tastes sweet and does not caramelize in heat is impossible because it would both include and not include sugar).

Table A.1. Potential Observations Resulting from 6 Experiments

	Visual	Water	Iodine	Vinegar	Heat	Taste
0	No Obs.	No Obs.	No Obs.	No Obs.	No Obs.	No Obs.
1	Crystal	Dissolve	Not Blue	No Fizz	Nothing	Tasteless
2	Powder	Gooey mess	Blue	Fizz	Brown	Sweet
3	Mixture	Lumpy/muddy			Caramelize	Salty
4		Lumpy/hardens				Sweet and Salty

Table SQL within the spreadsheet represents all possible accurate deductions for a given set of observations, conditional on particular minimum and maximum settings. Each line in the table includes a bracket containing seven fields, for example:

('322030', 3, '011000', '011X0X', '011101', 2, 4)

The seven fields are as follows:

1) observations from sets of selected experiments

2) decision rules for applying different minimums and maximums

3-5) three sets of deductions, based on the decision rules

6) break 1, the first threshold for applying the decision rules

7) break 2, the second threshold for applying the decision rules

The first field of 6 digits represents the following experiments: 1) visual, 2) water, 3) iodine, 4) vinegar, 5) heat, and 6) taste. The presence of a non-zero number in the corresponding field indicates that an experiment has been run. The particular non-zero values indicate the observations from a given experiment (see Table A.1). In our example, '322030' indicates:

- the visual experiment was run and indicated a mixture of powder and crystal
- the water experiment was run and the mixture turned into a gooey mess
- the iodine experiment was run and the mixture turned blue
- the vinegar experiment was not run
- the heat experiment was run and the mixture caramelized
- the taste experiment was not run

The second field indicates the use of decision rules based on different minimums and maximums. Some powder mixes are solvable given some sets of minimums and maximums, but not other sets. For instance, if the maximum number of powders was two and the taste experiment indicated sweet and salty, a solution could be reached (the mixture includes salt and sugar only); if the maximum number of powders was higher, we could not reach a solution. Four different decision rules are possible:

- A '1' indicates that the given deductions (field 3) apply in all cases (e.g., with all possible minimum and maximum combinations)
- A '2' indicates that the first listed deductions (field 3) apply for cases in which the maximum number of powders is less than or equal to a given threshold (break 1 in field 6); otherwise the second listed deductions (field 4) apply

- A ‘-2’ indicates that the first listed deductions (field 3) apply for cases in which the minimum number of powders is less than or equal to a given threshold (break 1 in field 6); otherwise the second listed deductions (field 4) apply
- A ‘3’ indicates that the first listed deductions (field 3) apply for cases in which the maximum number of powders is less than or equal to a given threshold (break 1 in field 6), the second listed deductions (field 4) apply when the maximum is greater than the first given threshold (break 1) and the minimum is less than a second given threshold (break 2 in field 7), and the third listed deductions (field 5) apply when the minimum is greater than or equal to the second given threshold (break 2)

In our example, the second field is a 3; two thresholds are provided and applied in the evaluation algorithm.

Fields 3, 4, and 5 (with 6 characters each) represent accurate and complete deductions that can be made following each experiment. In our example, these deductions strings are ‘011000,’ ‘011X0X,’ and ‘011101.’ The deductions represent what can possibly be deduced about the presence/absence of each powder in the mixture given the implemented experiments and associated results. Coded similarly to the final response string described previously, a ‘0’ indicates that a powder can be determined to be out of the mixture, a ‘1’ indicates that a powder can be determined to be in the mixture, an ‘X’ indicates that there is not enough information to determine the status of a powder (there are more experiments that could be run and provide more information), and an ‘N’ indicates that the status of a powder cannot be determined given all possible experimental results. In our above example, the deduction string ‘011000’ indicates that sugar and cornstarch are present and everything else is absent.

As a reminder, in our example:

(‘322030’, 3, ‘011000’, ‘011X0X’, ‘011101’, 2, 4)

the first field indicates that the visual, water, iodine, and heat experiments have been run with the following results: mixture, gooey mess, blue, and caramelize. Since the 3 in the second field indicates that deductions are dependent on specified minimums or maximums, we evaluate three cases using the thresholds in fields 6 and 7.

Case 1 deduction: ‘011000’ when the given maximum is less than or equal to 2 (break 1 in field 6).

In this case, sugar and cornstarch are present and everything else is absent. Given a maximum of 2 substances, a mixture observation indicates that one crystal and one powder are present. Table A.2 indicates the substances that can be inferred given particular observations and selected experiments; in this case, cornstarch in water results in a gooey mess (this would not be obscured by sugar or salt), and blue in vinegar indicates the presence of cornstarch (again). Finally, caramelization only occurs with the presence of sugar (4^{th} column, Table A.2). Therefore, sugar and cornstarch are the 2 substances.

Table A.2. Observations with Combinations of Experiments and Substances

	Visual	Water	Heat	Taste	Iodine	Vinegar
Flour	Powder	Lumpy/muddy	Brown	Tasteless	Not Blue	No Fizz
Sugar	Crystal	Dissolve	Caramelize	Sweet	Not Blue	No Fizz
Cornstarch	Powder	Gooey mess	Brown	Tasteless	Blue	No Fizz
Soda	Powder	Dissolve	Nothing	Tasteless	Not Blue	Fizz
Plaster	Powder	Lumpy/hardens	Nothing	Tasteless	Not Blue	No Fizz
Salt	Crystal	Dissolve	Nothing	Salty	Not Blue	No Fizz

Case 2 deduction: '011X0X' (field 4) when the given maximum is greater than 2 (field 6) and the given minimum is less than 4 (field 7). The only possible number of powders is 3.

In this case, sugar and cornstarch are present, flour and plaster are absent, and soda and salt are undetermined. Based on the same logic as Case 1, cornstarch and sugar are present. The gooey mess indicates the absence of plaster and flour (see Table A.2). However, no test results help distinguish the third powder as soda or salt.

Case 3 deduction: '011101' (field 5) when the given minimum is greater than or equal to 4 (field 7).

Based on the logic in Cases 1 and 2 and given a minimum of four substances, soda and salt must be present because plaster and flour are not present.

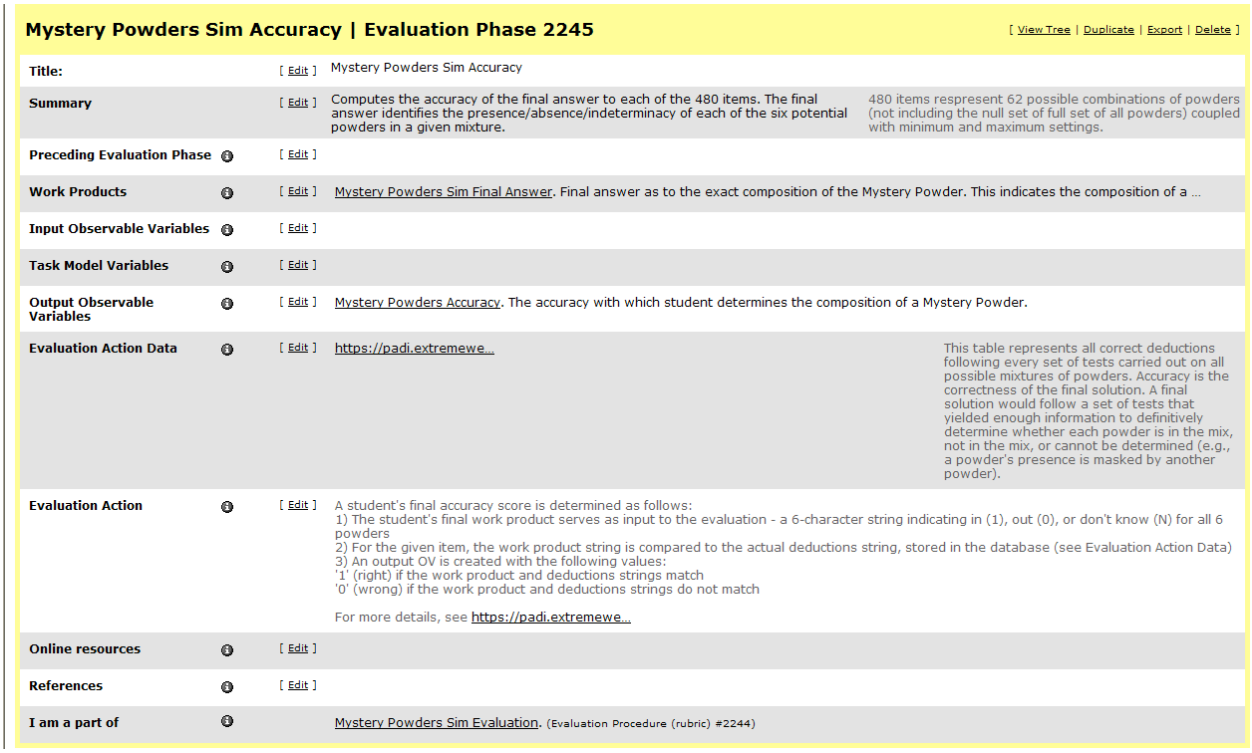
For a given task, a simple comparison of the "Optimal" Decision string with the Examinee Decision string (Work Product) yields our measure of Accuracy.

(3) Output OV Created from Evaluation Algorithm

An output OV is created with the following values:
1 = right: the Examinee Decision (Work Product) and Optimal Decision strings match
0 = wrong: the Examinee Decision and Optimal Decision strings do not match

Figure A.1 provides the designer's view of the Evaluation Phase for Accuracy in the Mystery Powders Simulation *template*.

Figure A.1. Evaluation Phase for Accuracy in the Mystery Powders Simulation template.



Mystery Powders Sim Accuracy | Evaluation Phase 2245 [View Tree | Duplicate | Export | Delete]

Title:	[Edit]	Mystery Powders Sim Accuracy
Summary	[Edit]	Computes the accuracy of the final answer to each of the 480 items. The final answer identifies the presence/absence/indeterminacy of each of the six potential powders in a given mixture. 480 items respresent 62 possible combinations of powders (not including the null set of full set of all powders) coupled with minimum and maximum settings.
Preceding Evaluation Phase	[Edit]	
Work Products	[Edit]	Mystery Powders Sim Final Answer. Final answer as to the exact composition of the Mystery Powder. This indicates the composition of a ...
Input Observable Variables	[Edit]	
Task Model Variables	[Edit]	
Output Observable Variables	[Edit]	Mystery Powders Accuracy. The accuracy with which student determines the composition of a Mystery Powder.
Evaluation Action Data	[Edit]	https://padi.extremewe... This table represents all correct deductions following every set of tests carried out on all possible mixtures of powders. Accuracy is the correctness of the final solution. A final solution would follow a set of tests that yielded enough information to definitively determine whether each powder is in the mix, not in the mix, or cannot be determined (e.g., a powder's presence is masked by another powder).
Evaluation Action	[Edit]	A student's final accuracy score is determined as follows: 1) The student's final work product serves as input to the evaluation - a 6-character string indicating in (1), out (0), or don't know (N) for all 6 powders 2) For the given item, the work product string is compared to the actual deductions string, stored in the database (see Evaluation Action Data) 3) An output OV is created with the following values: '1' (right) if the work product and deductions strings match '0' (wrong) if the work product and deductions strings do not match For more details, see https://padi.extremewe...
Online resources	[Edit]	
References	[Edit]	
I am a part of		Mystery Powders Sim Evaluation. (Evaluation Procedure (rubric) #2244)

APPENDIX B

Appendix B: Evaluation Procedures to Determine the Acuity Observable Variables

B.1 Summary of Logic for Determining the Acuity Stepwise OVs and Final OV

1) Since stepwise (and final) Acuity OVs are based on non-final steps within a Mystery Powders task, if the examinee gives a final answer after only one experiment, no Acuity OVs are determined.

2) Stepwise Acuity OVs are calculated individually before being combined into one final Acuity OV.

3) All stepwise Acuity OVs are determined as follows:

- The input to the stepwise Acuity Evaluation Phase is an examinee's Work Product of deductions following the *first* non-final experiment with the powder mixture—that is, a 6-character string indicating in (1), out (0), or don't know (X) for all 6 powders (see Appendix A).
- This work product (deduction) string is compared to the actual deductions string, stored in the database (pointed to in the template in the Evaluation Action Data object).
- Based on the percentage of correct deductions, a Step 1 Acuity OV is created with these possible values:

  100% for 6 (of 6) correct deductions

  83% for 5 (of 6) correct deductions

  67% for 4 (of 6) correct deductions

  50% for 3 (of 6) correct deductions

  33% for 2 (of 6) correct deductions

  17% for 1 (of 6) correct deductions

  0% for 0 (of 6) correct deductions

- This procedure is repeated for all non-final steps. For example, if the examinee provided a final solution to the task after 5 experiments, this stepwise Evaluation Phase would be carried out 4 times, creating 4 stepwise OVs.

4) A final Acuity Evaluation Phase is carried out, combining the stepwise Acuity OVs and resulting in one final Acuity OV. In this phase, the stepwise OVs are averaged; the average is then scored according to the following percentage ranges:

Score	Percentage Range (Average)
0	Up to 60%
1	61% - 80%
2	81% - 99%
3	100%

Thus, one final Acuity OV is created with a value of 0, 1, 2, or 3.

Work Product Input to the Stepwise Evaluation Algorithm

For any given Mystery Powders task, following each experiment with the mixture, the examinee decides whether each of the six potential powders is in, out, or unknown, yielding a set of 6 decision responses (see Appendix A). Non-final responses are evaluated and contribute to the Acuity OV; final responses are evaluated and contribute to the Accuracy OV.

Evaluation Algorithm: Comparison of Examinee and Optimal Decisions

Accuracy is the quality of deductions for an examinee's stepwise solutions to a Mystery Powders task. Determination of Acuity involves a table within a spreadsheet named 'deductions_2FIX.xls' (linked to in the PADI object model as Evaluation Action Data). The SQL table within the spreadsheet provides all possible accurate deductions for every set of experiments and results, given all possible minimums and maximums. Table A.1 in Appendix A indicates the possible observations for each of the 6 experiments that can be run.

A description and example of how examinee and optimal decisions are compared is provided in Appendix A.

For a given non-final experiment within a task, the Examinee Decision string is compared with the "Optimal" Decision string. The percentage of correct deductions is the stepwise Acuity OV; these are later combined and scored for a final Acuity OV.

APPENDIX C

Appendix C: Evaluation Procedures to Determine the Efficiency Observable Variables

C.1 Summary of Logic for Determining the Efficiency Stepwise OVs and Final OV

1) Stepwise Efficiency OVs are calculated individually before being combined into one final Efficiency OV.

2) The first stepwise Efficiency OVs is determined as follows:

- The input to the stepwise Efficiency Evaluation Phase is an examinee's first choice of experiment, given a specified minimum and maximum numbers of powders.
- This first choice of experiment is compared to the optimal choice of experiment, retrieved from the database (pointed to in the Evaluation Action Data object; see details below).
- If the first choice of experiment is optimal (a value of 1.00), the Step 1 Efficiency OV is created with a score of 1. Otherwise it is given a score of 0.
- This procedure is repeated for all steps within the task. Subsequent steps take previous experiment choices and resulting observations into account (see below). Stepwise OVs are created in this Evaluation Phase corresponding to each choice of experiment.

3) A final Efficiency Evaluation Phase is carried out combining the stepwise Efficiency OVs and resulting in one final Efficiency OV. If all of the stepwise OVs have a value of 1, the final Efficiency OV is scored as 1; otherwise, the final Efficiency OV is scored as 0. Thus, one final Efficiency OV is created with a value of 0 or 1.

C.2 Evaluation Algorithm: Determination of Stepwise Efficiency

Inputs to the stepwise efficiency evaluation algorithm include previous experimental results (from previous steps) and the examinee's selection of a new experiment. Also needed are the minimum and maximum settings for a given task.

Determination of stepwise Efficiency involves the 'optimals_for_gen.xls' database. The spreadsheet table 'optimals_for_gen' represents the efficiency of all possible choices of experiments given each set of minimums and maximums. The first column in this spreadsheet is a 12–digit field representing a current set of experiments and results as well as the previous sets of experiments and results (from the previous step). The next 15 columns represent all possible combinations of minimums and maximums: (1,1), (1,2), (1,3), (1,4), (1,5), (2,2), (2,3), (2,4), (2,5), (3,3), (3,4), (3,5), (4,4), (4,5), (5,5); the corresponding table entries represent the Efficiency of choosing a particular experiment, given a particular minimum and maximum. Efficiency ratings range from 0 to 1; negative entries indicate impossible combinations. An example (which happens to be line 51 in the aforementioned Evaluation Action Data file) in which the first field is:

000023001011

The first 6 digits represent the previously run experiments and observational results available at the point choosing a subsequent experiment. The digits represent, from left to right, the following experiments: 1) visual, 2) water, 3) iodine, 4) vinegar, 5) heat, and 6) taste. In our example, the string '000023' indicates that the heat and taste experiments have been run, that the heat experiment resulted in a brown mixture, and that the powder mixture tasted salty (see Table C.1 for observation values). The second set of six digits represent the already completed experiments as well as the newly selected experiment (0 = not chosen, 1 = chosen). In our case, '001011' indicates the iodine, heat, and taste experiments. Because the first string offers results for the heat and taste experiments, we can conclude that the newly selected experiment is the iodine test.

Table C.1. Potential Observations Resulting from 6 Experiments

	Visual	Water	Iodine	Vinegar	Heat	Taste
0	No Obs.	No Obs.	No Obs.	No Obs.	No Obs.	No Obs.
1	Crystal	Dissolve	Not Blue	No Fizz	Nothing	Tasteless
2	Powder	Gooey mess	Blue	Fizz	Brown	Sweet
3	Mixture	Lumpy/muddy			Caramelize	Salty
4		Lumpy/hardens				Sweet&Salty

The following 15 fields refer to the Efficiency of the selected experiment, in our case the iodine test, given different sets of minimums and maximums. Each field is named according to the minimum and maximum number of powders that a student has been told can appear in his mystery powder; e.g., column s13 refers to the case in which the student is told there are at least one and at most three powders in his mixture. These Efficiencies take into account the previously run experiments. Without previously run experiments selected, the Efficiency of a newly selected experiment will only depend on the minimum and maximum settings. Lines 2-7 in the 'optimals_for_gen.xls' spreadsheet represent these cases of no prior experiments. Each line represents the selection of a different experiment (e.g., line 2 is taste, line 3 is heat). Efficiency ratings range from 0 to 1. Lines 2-7 in the file indicate that the taste experiment is generally the most efficient (efficiency = 1.00), except for cases with lower maximums: Efficiency = 0.60 for (1,1), 0.80 for (1,2), and 0.80 for (2,2). In these cases, water is the most Efficient experiment. It should be noted that Efficiencies are rank-ordered and evenly spaced across all possible experiment choices for a given scenario. In the case of no prior experiments, the available 6 experiments are ranked in Efficiency at levels 0.00, 0.20, 0.40, 0.60, 0.80, or 1.00 for each pairing of minimums and maximums. In the case of the two prior experiments of heat and taste, efficiency ratings are 0.00, 0.33, 0.67, or 1.00.

Back to our example (line 51) in which the prior heat experiment resulted in brown, the mixture tasted salty, and the iodine experiment was selected. The -1 in column s11 (minimum and maximum of 1) indicates that this scenario is considered impossible; this is because the salty taste indicates the presence of salt, and the turning brown indicates the presence of flour or cornstarch. Therefore, at least two powders must be present. The 1.00 in columns s12, s22, and s44 indicates that the iodine experiment is the most

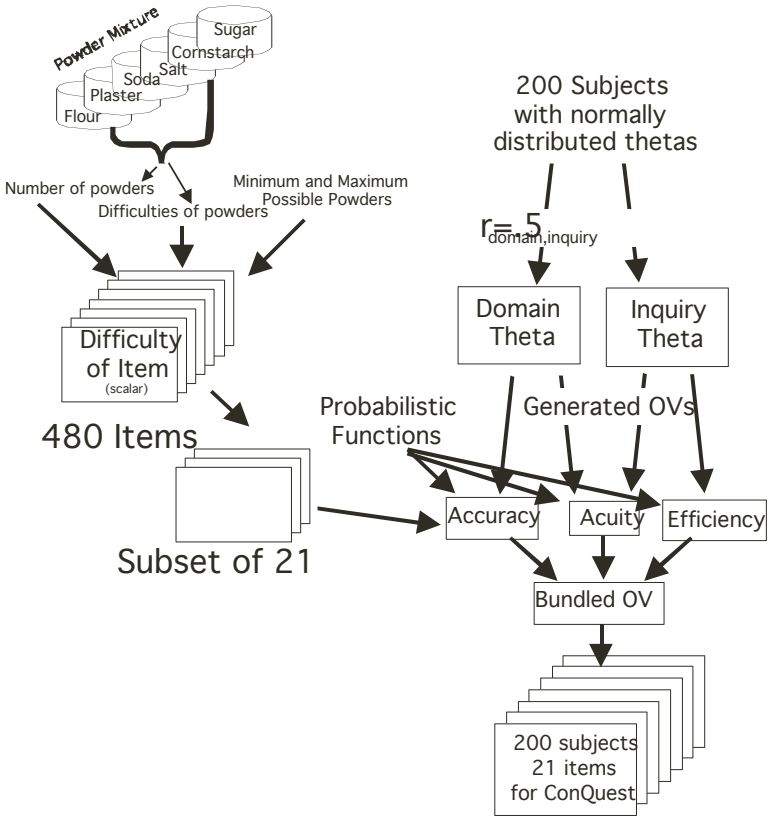
efficient for these particular sets of minimums and maximums (since it is likely to differentiate between flour and cornstarch. The iodine experiment is rated 0.67 in efficiency for columns s13, s23, and s33; in these cases, water is considered a more efficient experiment since it will either differentiate between flour and cornstarch *or* indicate the presence of plaster (more information than the iodine experiment). The iodine test is rated 0.67 in efficiency for column s45; in this case, vinegar is considered a more efficient experiment than iodine. The iodine experiment is rated 0.33 in efficiency for columns s14, s15, s24, s25, s34, and s35; in these cases, water is considered the most efficient experiment (1.00), and vinegar is considered a more efficient experiment (0.67) than iodine (probably because vinegar will definitively identify the presence of soda – something there is currently no information about - and we already have some information about cornstarch).

APPENDIX D

Appendix D: Generating Simulated Data for Task Calibration

The program for simulating response data was an Excel macro, written in Visual Basic, that took advantage of the programmed logic present in the mockup. Figure D.1 shows the flow of the generation process.

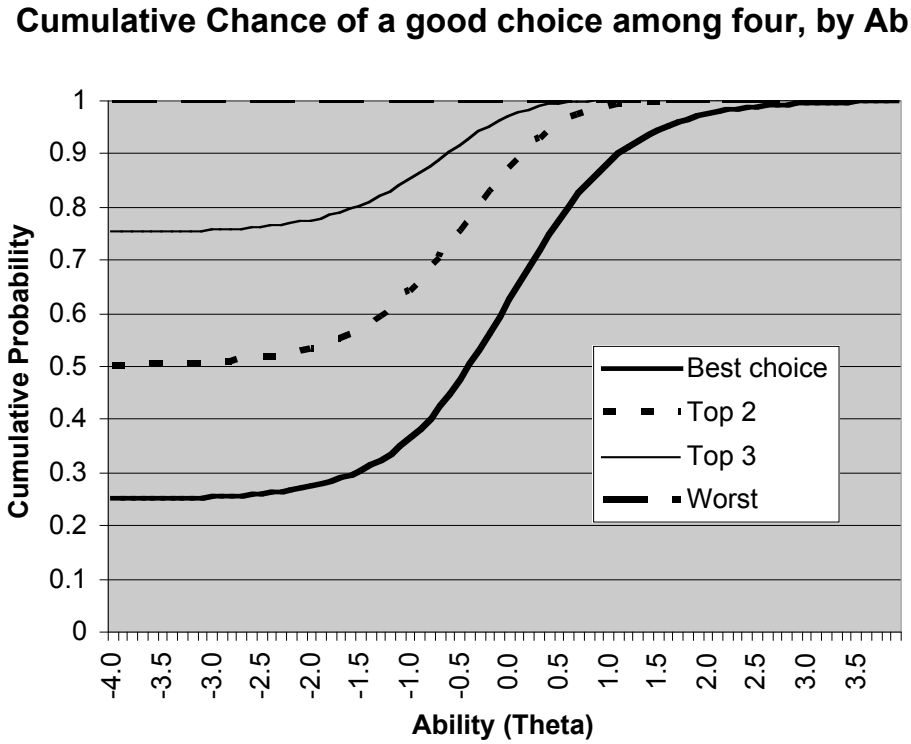
Figure D.1. Generation of Simulated Examinee Data for ConQuest.



We generated values for two Student Model Variables (SMVs), inquiry ability and domain knowledge, for 200 simulees, with unit-normal distributions with a correlation of .5. We assumed the difficulty of tasks would depend on three task variables (as in Embretson, 1998, Bejar, 2002, and Hornke & Habon, 1986).

We then generated three scores (OVs) for each simulee-task combination: Accuracy, Acuity, and Efficiency. Recall that Accuracy is a function only of domain knowledge, Efficiency only a function of inquiry ability, and Acuity a function of both. We used probabilistic models to simulate the likelihood of a correct answer on Accuracy and Efficiency, which are scored dichotomously. For Acuity, we probabilistically generated four levels of response based on simulee ability where Ability in this simulation was defined as the average of the two SMVs, inquiry ability and domain knowledge. Figure D.2 shows the distribution that was used.

Figure D.2. Rasch Logic for Probability of Optimal Choice among Multiple Levels as a Function of Ability



The two dichotomous and single four-level OVs were combined into the single 16-level bundled variable (2x2x4 = 16) for each simulee-task combination. The ConQuest computer program (Wu, Adams, & Wilson, 1998) was used to estimate parameters for task, which were then added into the Measurement Model section of the *template* for use in the test scoring process.

References

Bejar, I.I. (2002). Generative testing: From conception to implementation. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Hillsdale, NJ: Erlbaum

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

Hornke, L. F., & Habon, M. W. (1986). Rule-based bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement, 10,* 369-380.

Wu, M., Adams, R. J., & Wilson, M. R. (1998). *ConQuest* [computer program]. Camberwell, Victoria: Australian Council for Educational Research.