At this symposium, assessment experts, curriculum developers, and software engineers from the Principled Assessment Designs in Inquiry (PADI) Project will present and demonstrate the online assessment design system and supporting resources as they are instantiated in three national science inquiry curricula. The PADI project is comprised of an interdisciplinary, highly collaborative research and development team from SRI International, University of Maryland, University of California, Berkeley, The Lawrence Hall of Science and the University of Michigan. In its fourth year of funding from the Interagency Education Research Initiative (IERI), this project has developed (1) an online assessment design system that is based on evidence-centered design (ECD) in the realm of science assessment for inquiry (Mislevy, Sternberg, & Almond, 2002); (2) a collection of assessment task templates reverse-engineered from widely-used assessment tasks (i.e., NAEP Floating Pencil (O’Sullivan, Reese, & Mazzeo, 1997), Mystery Boxes (Baxter, Elder, & Glaser, 1996) and Mystery Powders (Baxter, Elder, & Glaser, 1995) to show proof-of-concept that the PADI Design System can accommodate complex assessments of inquiry; (3) a collection of assessment design patterns, templates, and tasks built around the complex-problem solving that is the goal of three of the nation’s most widely implemented science inquiry curricula—BioKIDS, FOSS, and GLOBE; and (4) an additional set of technology resources to facilitate the assessment design and implementation processes. These resources include a scoring engine based on Item Response Theory (IRT) that accommodates a family of multivariate psychometric models (Kennedy, in press); a calibration engine that estimates the item parameters of the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Adams, Wilson & Wang, 1997); wizards (web forms) that scaffold the design process; and Gradebook, an application that manages student data in conjunction with the PADI design system and the IRT-based scoring engine (Hamel, in press).

In this symposium, AERA members will explore the technology that allows assessment developers to employ the expertise embedded in the PADI system to implement assessment designs for inquiry-based science instruction. However, the PADI tools and resources are eclectic and can guide assessment development in any subject area, based on any psychological tradition (e.g., cognitivist, socio-cultural), and using any item format.

The symposium objectives are to:
1. Illustrate how the use of the PADI project’s conceptual framework and online tools can transform assessment development from a craft toward the science of design

2. Illustrate how the PADI design system supported the incorporation of principles of ECD in the design of inquiry assessments for three widely-implemented science curricula

3. Illustrate how the PADI design system and related resources support the use of complex measurement models that can meet the goals of assessments with varied purposes

4. Provide opportunities for symposium participants to explore the PADI system at activity stations and interact with curriculum developers, psychometricians, software engineers and assessment developers about the system

**Significance**

The current pressures from the No Child Left Behind legislation to measure students' academic performances and from assessment research calling for increased assessment quality demand new approaches for the assessment of students’ learning. Educators are called on to develop new large-scale assessments and new classroom assessments that can be used diagnostically to improve student outcomes. To address these needs, the PADI project has incorporated the tenets of evidence-centered design in an online system that facilitates the development of the types of high-quality assessments advanced in *Knowing What Students Know* (National Research Council, 2002).

**Symposium Structure**

This interactive symposium will have three parts: (1) two opening presentations that will overview PADI, (2) an interactive segment featuring four illustrative ‘stations’, and (3) a final discussion led by two discussants—Dr. John Behrens (Cisco Systems) and Dr. Edward Haertel (Stanford University). Dr. Janice Earle (National Science Foundation) will serve as Chairperson.

**1. Overview Presentations (30 minutes)**

Presentation 1: Overview of PADI Concepts and Components

<table>
<thead>
<tr>
<th>Robert Mislevy</th>
<th>Geneva Haertel</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Maryland</td>
<td>SRI International</td>
</tr>
</tbody>
</table>
This presentation overviews the theoretical foundations of PADI (including evidence-centered design) and how they are implemented in the PADI design system and its supporting technological resources. Assessment designs are based on tacit conceptions that have evolved over the last century. Many assessment design processes and artifacts may seem simple because they are familiar, but the underlying principles are not. As designers define new, more complex kinds of assessments (e.g., problem solving in simulation environments) or work with different kinds of data (e.g., traces of actions in open-ended cycles of inquiry), the complexity of design principles becomes clear (NRC, 2002). In these cases, it is productive to identify at a higher level of generality the elements and relationships that are common among successful assessments of all types, leveraging unifying concepts and relationships. Drawing on Messick’s (1994) seminal work on ECD, the PADI design system lays out the essential elements of a coherent assessment argument (i.e., the student, evidence, and task models) and makes explicit the ‘layers’ of associated design decisions (Riconscente and Mislevy, 2005).

In PADI, technology links the conceptual and psychometric elements of the design process. Using Web-based forms, assessment designers construct the assessment argument, first in narrative form (i.e., design patterns), and then by specifying the details of the student, evidence, and task models (i.e., task templates and task specifications) (Mislevy et al, 2003; Riconscente & Mislevy, 2005). These Web-based guiding structures are expressed as extensible object models. Object modeling, an approach to software design, helps the assessment designer address critical considerations of consistency, usability, and validity. In addition, object modeling resolves issues of limited replicability, scalability, and cost-effectiveness, which have plagued many previous efforts to design complex assessments efficiently in meaningful contexts.

The Scoring and Calibration Engines developed for the PADI project facilitate realistic testing of the PADI Design System. The scoring engine enables modeling of complex multivariate assessment tasks, including those with varied item formats and those involving sequential or interdependent responses. The Calibration Engine is used in conjunction with Conquest (ACER, 2005) to investigate model fit and establish item and model parameter estimates for instantiated tasks. The Scoring Engine is used in conjunction with the PADI Gradebook application to produce expected a-posteriori, plausible value, or maximum likelihood estimates of proficiency. Together, these PADI-compatible engines close the loop from assessment design to gathering relevant evidence and drawing inferences consistent with the assessment purpose.

Presentation 2: Overview of PADI Instantiations

| Cathleen Kennedy | Larry Hamel |
The second introductory presentation will review how each curriculum group developed design patterns and then specified the assessment argument in task templates and task specifications. Discussing the design process as it was instantiated by different PADI curriculum partners will highlight the important features of PADI and how the system can be used in multiple contexts. An example of the design-to-implementation process will be discussed and graphically illustrated making explicit the interaction and sequence of the components of design process and the PADI design system. This presentation will prepare audience members for the interactive segment of the symposium by providing necessary background; discussion at activity stations can then focus on deeper issues relating to participants’ area of interest.

2. Activity Stations (30 minutes)

In the interactive phase of the symposium, four ‘stations’ will receive audience members. At each station, PADI team members will further describe the conceptual, technical and technological accomplishments of PADI as implemented in a particular curriculum context or through the reverse-engineering of a performance assessment. All stations will have laptops so that audience members can view the functioning Web-based PADI design system. Visual representations of the design and implementation process will be presented for each station. Thus, the audience will come to understand the reusability and scalability of the PADI system by seeing its instantiation in multiple curricular contexts and in reverse-engineered assessment tasks. PADI team members will be at each station to interact with audience members about their area of expertise (see Table of participants below).

<table>
<thead>
<tr>
<th>Expertise at Activity Stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOSS Station</td>
</tr>
<tr>
<td>Assessment Expertise</td>
</tr>
</tbody>
</table>
### A. Foss Activity Station

**FOSS Background.** FOSS (Full Option Science System) has developed middle school courses in three science content areas: earth/space, life, and physical sciences/technology (http://www.lhs.berkeley.edu/FOSS). The FOSS curriculum sees scientific inquiry as seeking answers to questions about the natural world and centers on three important constructs: science content, conducting investigations, and building explanations. FOSS differentiates measures of inquiry from content knowledge.
Instantiation into PADI design system. Aspects of inquiry that were important to measure for diagnostic and accountability purposes and that reflected the NSES Inquiry Standards were incorporated into the ASK Principles of Scientific Inquiry Design Pattern. An analysis of the KSA (Knowledge, Skills, and Abilities) that were relevant to the constructs being measured determined that both assessment purposes could be accommodated using the same assessment tasks. Two Student Models, one for each purpose, and three types of assessment activities (Performance Tasks, Inquiry Scenarios, and I-Checks (quiz-type benchmark assessments)) were created. FOSS also created three Activity objects that describe the layout and measurable components of these activities. Finally an assessment Template was constructed that put all the pieces together and generated Task Specifications that linked specific Student Models and each of the Activity types.

Lessons Learned. PADI directed the analysis and description of the FOSS assessment context and components. Completing the Design Pattern crystallized FOSS’ thinking about the purpose of its assessments. Creating Activity structures and Student Models ensured that appropriate evidence about students’ inquiry abilities was consistently elicited. As a result of this experience, FOSS is currently developing an item authoring wizard that asks authors to generate a set of Task Specifications that lay out the structure of new assessment items and identify how their items elicit required evidence in these multiple assessment contexts.

B. BioKIDS Activity Station

BioKIDS Background. The BioKIDS: Kids’ Inquiry of Diverse Species Project (Songer, 2000) focuses on developing, implementing and evaluating inquiry-based, technology-rich science curricula and assessments in high-poverty, urban areas.

Instantiation into PADI Design System. The BioKIDS team identified key aspects of scientific reasoning embedded in the BioKIDS project such as building scientific explanations. Like the FOSS project, they used PADI design patterns to translate these aspects of inquiry into assessment arguments. In addition, they created an activity matrix where one dimension described content complexity and another described the degree of support or scaffolding provided to students. Once these were in place, they authored tasks in accordance with their matrix and task specifications, administered an assessment to a sample of students, and fit the data with the MRCMLM. The MRCMLM determined (1) the dimensionality of their assessment items regarding the measurement of both scientific content and inquiry skills, (2) the difficulty of the items, comparing empirical to hypothesized difficulty given the placement of the items on their
matrix, and (3) the nature of students’ knowledge before and after participation in the curriculum.

**Lessons Learned.** One important result of the BioKIDS analyses, which used test data with multi-part responses and a mixture of content and inquiry demands, was that a two-dimensional model of student understanding was able to generate scores that captured student achievements on both an overall inquiry and content-specific dimensions (Songer et al., 2005). This result confirms that the PADI design system will enable future developers to generate assessments tasks based on hypothesized student models that can be subsequently confirmed or disconfirmed using empirical findings.

**C. GLOBE Activity Station**

**GLOBE Background.** GLOBE is a “worldwide hands-on, primary and secondary science program” (http://www.globe.gov) where students take measurements related to earth science content investigation areas.

**Instantiation into PADI Design System.** The GLOBE team developed classroom assessment tools including an assessment framework, a template to guide teachers’ development of classroom assessments, and sample and generic scoring rubrics. Classroom assessment resources were reverse-engineered into PADI design system components. The GLOBE team developed several assessment design patterns, a complex task template, and a web-Wizard to assist assessment developers with generating task specifications related to specialized GLOBE assessments. Simulated data was also used to test and refine the measurement models.

**Lessons Learned.** GLOBE is an important use-case demonstrating the capability of PADI to accommodate assessments that address multiple components of science inquiry such as planning, data analysis, and communication. Thus, the challenges and lessons learned can be generalized to support the design of similar complex assessments of science inquiry. Of particular interest are differences in representing multiple phases of inquiry in design patterns as compared to task templates, the value of including multiple student models that vary in specificity within a task template, strategies for implementing task model variables such as scaffolding, and how to address item dependencies.

**D. Performance Task Activity Station**

This station presents three complex and well-known performance assessments that were reverse-engineered using the PADI Design System. The reverse-engineering process afforded opportunities to explicate the assessment logic underlying these tasks, to identify potential extensions or modifications to the
original task, and to explore the strengths and weaknesses of the PADI Design System.

The Floating Pencil Task. The Floating Pencil task is a science performance assessment administered as part of the 1996 NAEP. The 8th grade version consisted of 14 multiple-choice and open-ended items. Students conduct a hands-on investigation using standardized laboratory materials in which the experimental procedures have been specified. A ‘family of tasks’ of which the Floating Pencil task is a member was identified and led to the development of an abstract template that could generate other similar tasks. The PADI team reverse engineered the task model (e.g., materials, work products), evidence model (evaluation procedures, statistical models), and student models to develop a Task Specification.

The Mystery Powders Task. This assessment (Baxter, Elder, & Glaser, 1995) engages students in systematic investigation of the properties of six white powders (e.g., salt, sugar). Each powder has a unique set of physical and chemical properties that can be determined by visual inspection, taste, application of heat, or chemical reagent. Students select appropriate tests, analyze the simulated results, and conclude which powders are present in the Mystery Powder. The task was varied to be suitable for Internet delivery by using the PADI four-process architecture (Item selection, Presentation, Answer evaluation, Student model update) to deliver the assessment via QTI (IMS Consortium, 2000). Thus, it is possible to deliver 30 assessment items (powder combinations) in an hour, as opposed to one in the original experiment.

The Mystery Boxes Task. This task (Baxter, Elder, & Glaser, 1996), assesses 5th students’ content knowledge about electric circuits. Students use wires, bulbs, and batteries to identify each of six box’s circuit-related contents. Reverse-engineering began with the drafting a Design Pattern. The design and measurement consequences of different measurement models were examined and templates were created to express the potential variation. Twenty objects were created to represent the assessment argument for the problem-solving demands.

3. Discussion (30 minutes)

The discussants, Dr. John Behrens (Cisco Systems) and Dr. Edward Haertel (Stanford University) will comment on the project’s progress to date and issues emerging from the session. Questions from the audience will be addressed.