Using the PADI Design System to Examine the Features of a NAEP Performance Assessment

Prepared for AERA
April 12, 2005
Montreal

Kathleen C. Haynie, Kathleen Haynie Consulting
Andrea A. Lash, SRI International
Geneva D. Haertel, SRI International
Edys S. Quellmalz, SRI International
Angela Haydel DeBarger, SRI International

Using the PADI Design System to
Examine the Features of a NAEP Performance Assessment

Introduction

As we enter the 21st century, use of collaborative workplace tools is on the rise. In the field of education, such socio-technical tools facilitate the development of previously unimaginable educational resources. Through support from NSF, the Principled Assessment Design for Inquiry (PADI) project seeks to improve the assessment of inquiry in science learning; this effort is a multi-organization collaboration, involving the development and use of a Web-based data modeling tool.

Collaborative practice involving computer-based tools has recently become an important field of study. The field of Human-Computer Interaction (CHI) is concerned with interactive computing systems for human use and the major phenomena surrounding them and has been studied extensively since the mid-1980's. Two pertinent branches of CHI are Computer-Supported Cooperative Work (CSCW) and Computer-Supported Collaborative Learning (CSCL). The Computer-Supported Cooperative Work (CSCW) involves the design and use of technologies that affect groups, organizations, communities, and societies; these typically involve business settings, with a focus on communication and productivity. Computer-Supported Collaborative Learning (CSCL) involves the study of how students are scaffolded or supported in learning together effectively. Theoretical grounding for this work is provided through Vygotsky's sociocultural theory, constructivist principles, distributed cognition (e.g., Brown, Ash, et al., 1993), and situated cognition (e.g., Greeno, Collins, & Resnick, 1996; Greeno, 1999; Greeno & T.M. Group, 1997).

The success of collaboration using technological workplace tools involves a number of factors. Gee (2000) notes that the new capitalism has given rise to new forms of affiliation and the importance of socio-technological design knowledge. An important form of affiliation is typically that of communities of practice organized around a common endeavor. Members of such communities bring extensive knowledge and expertise, knowledge is distributed (spread across various members and tools / technology), and knowledge is dispersed (networked across different sites and institutions). Olson & Olson (2000) suggest that for technologically rich "dispersed" collaboration to be successful requires: (1) common ground (shared understandings of the common endeavor), (2) the coupling of work (carefully crafted relationships between individual and joint work), (3) collaborative readiness: individual collaboration skills, leadership ability to design and resource communities of practice, and (4) readiness of collaborative technology (development of appropriate "groupware" and participants' abilities to access and use available resources (Luff, Heath, et al., 2003)). Deviation from any of these qualities creates strain on the relationships among teammates and requires changes in the work or processes of collaboration to succeed.

The PADI Project

The Principled Assessment Design for Inquiry (PADI) project aims to provide a practical, theory-based approach to developing high-quality assessments of science inquiry (Mislevy, Hamel, Fried, Gaffney, Haertel, Hafter, et al., 2003) by developing a rigorous design framework. PADI is a special-case implementation of the evidence-centered design (ECD) framework developed by Mislevy, Steinberg, and Almond (2002). The ECD framework is based on a construct-centered approach to assessment (e.g., Messick, 1994) and describes the components of assessment design: a Student Model, an Evidence Model, and a Task Model (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2003).

For a given assessment, the Student Model addresses the question of what complex of knowledge, skills, or other attributes should be assessed. Student Model Variables are the underlying constructs an assessment is designed to assess. These constructs may be based on any theory of learning, e.g., behaviorism, cognitive psychology, constructivism, situated cognition.

The Evidence Model addresses the question of what student behaviors or performances is expected to reveal those constructs. The Evidence Model lays out the argument for why and how the observations from a given task constitute evidence of Student Model Variables. The Evidence Model includes the Evaluative Sub-Model and the Statistical Sub-Model. The Evaluative Sub-Model provides the rules for evaluating evidence (e.g., student Work Products), which results in observable variables (e.g., a circled letter 'A' on a multiple-choice item is evaluated as 'correct' and given a score of '1'). The Statistical Sub-Model updates the Student Model using the Observable Variables (e.g., a score of '1' on a biology test contributes to a higher estimate of a student's score on biology knowledge).

The Task Model addresses the question of what tasks or situations should elicit certain examinee behaviors. A Task Model provides a framework for describing the situation in which examinees act; such environmental specifications can include instructions, tools, lab materials, and characteristics of stimulus material. The Task Model also includes specifications for the form of an examinee’s response (e.g., writing an essay, graphing data). A task describes circumstances meant to elicit information about what an examinee knows or can do.

A primary deliverable of the PADI project is a model of the domain of scientific inquiry assessment in the form of a web-based tool. This web-based model is known as the Example-based Modeling, or EMo, system (Schank & Hamel, 2004). A secondary deliverable is a library of examples of assessment blueprints. Both the EMo system and the library of examples are developed collaboratively. The PADI Design System, which incorporates EMo and the library of examples, is not a specific science assessment task or application, but is rather a framework, based on the domain model and ECD framework, that serves many functions. The framework serves as a development tool and library of test “blueprints” from which science assessments might be developed. The object-oriented Design System is structured around representational forms, representations (or examples), and system tools.

Representational forms are comprised of interchangeable objects, such as Work Products and Student Model Variables. A Design Pattern is a representational form that calls for the laying out of an assessment argument (explicating the relationship between Student, Evidence, and Task Models) in a narrative form, as well as the characteristic and variable task features of some ‘family’ of tasks. A Template is a representational form that requires more specific task-related information including a Student Model and Evidence Model. Task Specifications are specific instances of Templates; these are blueprints for individual tasks, expressed in the PADI framework.

The PADI library of examples is populated with representational forms, specified according to some real-world application. For example, one of the National Science Educational Standards (NSES) inquiry standards was developed into a Design Pattern called ‘Conduct

Investigation'. A second example is the development of a Template called 'Mystery Boxes' based on an already existing task with the same name. Both 'Conduct Investigations' and 'Mystery Boxes' representations could be used to generate any number of science assessment tasks.

The PADI system provides system tools for users. These include diagrams of interrelationships between PADI objects (according to the ECD framework), definitions of PADI objects via a glossary and embedded help buttons, and links to PADI technical reports.

The work of the PADI project is carried out through a networked, collaborative work team, in interaction with the PADI Design System. For the PADI team, collaboration involves development of the PADI Design System, development of assessment blueprint examples, and use of the PADI Design System for additional purposes (e.g., analyses of tasks). The PADI team is a comprised of expert psychometricians, cognitive scientists, software developers, science content specialists, assessment designers, and engineers; thus, expertise is distributed. Members of the PADI project are networked across different sites and institutions: SRI International, University of Maryland, UC Berkeley, Lawrence Hall of Science, and University of Michigan.

The quality of collaboration for the PADI project is very high. PADI members participate in a weekly team meeting that involves access to high-level experts (highly experienced team members), access to key knowledge such as updates on each PADI strand (a strand is a particular line of pursuit in connection with the PADI system), and an educational component in which strands take turns presenting their current work and receiving critical feedback from the PADI team. In this way, common ground is established, coupling of work is maintained, and everyone has access to needed knowledge and resources. The development of the PADI Design System was carried out collaboratively through this weekly meeting structure, and this has assured that the system well-models the domain of science assessment and is usable for all team members. PADI enjoys strong leadership monitoring, encouraging good collaborative practices; this helps provide common ground for all team members, collaborative readiness, and coupling of work. On the PADI team, leaders establish symmetry of status – everyone is given voice, and considered an expert on their respective domain. The leaders assure that the following practices are carried out: articulation of common goals, division of labor, use of complementary expertise, and joint ownership of the group processes and outcomes. The PADI system was designed to allow real-time access to the system, collaborative development of the object-oriented design system, objects with specific properties and relationships, and a graphic user interface that immediately reflects changes in PADI representations – facilitating remote collaboration.

The quality of PADI collaboration has improved since the project's inception. Schank and Hamel (2004) noted qualities of improved team collaboration including broader participation, more negotiation, increased validation, improved understanding, and elevated conversations. They made the following attributions for these improvements (Schank & Hamel, 2004). Use of examples led to broader participation and understanding across disciplines. The familiar Web interface led to broader participation and more negotiation. Remote access via the Web led to more negotiation. Finally, lack of object-oriented verbiage led to broader participation and more elevated conversations.

The central process for strands involved with developmental work on the PADI Design System can be stated as follows: the interaction between their existing assessment materials (e.g., assessment tasks) and the Evidence-Centered Design framework is mediated by various tools, representational forms, and representations utilized by collaborative workgroups. The assumption here is that this is an epistemological process in which knowledge is created and owned, primarily, at the group level. This mediated interaction could be for the purpose of drafting blueprints for guiding the development of new science assessments or for the purpose of understanding the features of an already existing assessment (e.g., through reverse engineering an already existing task into a PADI representation). Purposes other than blueprint development and task exploration or analysis are also possible (e.g., aiding with an intelligent tutoring system).

A new strand was created for the purpose of reverse engineering science inquiry tasks from a large-scale, national assessment, NAEP. In doing that, the intention was to use the PADI Design System as an analytical tool for understanding the deep and surface characteristics of the chosen task. This paper will discuss the process by which our group accomplished its goals. In doing that, this paper will focus on two guiding questions: (1) what types of knowledge were created in the development of a PADI Task Specification based on an existing performance assessment task? and (2) what types of knowledge were created as the team analyzed a task using the PADI Design System?

Reverse Engineering and Analysis Process

Over the course of the last 10 months, the reverse engineering and analysis process for this new PADI strand moved through four sequential stages: (1) the initial conceptualization of our work, (2) the selection of a set of items, (3) the exploration of that set of items, and (4) the development of a Task Specification[1]. The initial conceptualization occurred from May through July, 2004. During this time, our group chose our central focus – a set of NAEP and other items

[1] PADI objects and representational forms will be capitalized.

from large-scale science assessments for reverse engineering and analysis. NAEP is considered a national benchmark in K-12 achievement, and NCLB currently requires states to administer 4th and 8th grade NAEP tests to some districts. Also, we understood that large-scale science assessments had a number of potential item sets that involved hands-on problem-solving for some phase or phases of scientific inquiry. We had available a group of about 60 potential items, with expert ratings of inquiry and content knowledge, as well as empirical data from 18 classes of students[2]. The initial goals were to reverse engineer and analyze a task or set of items. In reverse engineering, we intended to create a PADI representation (a Design Pattern, Template, or Task Specification); there was a sub-goal of developing group-level expertise with the PADI Design System. The intention of the analysis was to understand the deep and surface characteristics of a set of items; additionally, our team sought to gain a deeper understanding of large-scale assessments in terms of the ECD framework. Our approach for this strand of PADI work was collaborative – our team had seven members: a psychometrician, four educational psychologists, an assessment designer, and an engineer. One of our Co-PI's, Geneva Haertel, was the leader of this project strand; the other Co-PI, Robert Mislevy, frequently contributed to our team meetings (he is also the lead developer of the ECD framework). Our team spanned two sites: SRI International and a site in New Jersey. Our group communicated via email, periodic conference calls, and informal conversations. In addition, we joined the PADI project's weekly conference call meetings; these served a communicative as well as educational function for our group members.

The selection of the set of items occurred from July through August, 2004. Our team selected a NAEP performance assessment task called Floating Pencil, targeted to 8th grade

[2] These data were made available through the SRI study, '*Validities of Standards-Based Science Inquiry Assessments: Implementation Study*'. See Quellmalz, Haertel, et al. (2004), Quellmalz & Haydel (2003), and Quellmalz & Kreikemeier (2002).

science students. Following that, we entered an exploration phase of the Floating Pencil task from August through December, 2004. In exploring the task, a Design Pattern and a partial Template were drafted. The next stage was the development of a Task Specification for Floating Pencil from December 2004 through March 2005. During the selection, exploration, and analysis of the Floating Pencil task, our group grappled with defining and understanding the elements of the Evidence-Centered Design framework – the Student, Evidence, and Task Models. This epistemic process was prompted and aided by interactions with the PADI system: the requirements of the representation structures, the PADI systems tools (e.g., diagrams, glossary, help tools, exemplars of representations), and associated PADI technical reports.

The Student Model

Selecting a Task to Reverse Engineer

The set of PADI Design Patterns, in conjunction with the NSES inquiry standards, was used as a basis for selecting a task, based on cognitive and performance demands (e.g., content knowledge, inquiry skills). In doing this, our team grappled with the basis for selecting items. We questioned whether to use NAEP's developer's codes or the NSES inquiry standards. The role of the PADI Design Patterns was explored to see if they could include criteria for selecting items. We analyzed the existing set of PADI Design Patterns in terms of coverage of the NSES inquiry standards, in relationship to each other, and in terms of coverage of phases of inquiry. Our team discussed using standards as the basis for identifying a set of items to reverse engineer. The available pool of NAEP and other items was analyzed in terms of: usefulness for measuring inquiry, the NSES inquiry standards (particularly, standards B, F, and G), and natural item groupings or theme blocks (e.g., items linked to common stimulus or common topic). We selected a set of items – the Floating Pencil performance assessment task – for analysis and reverse engineering. The Floating Pencil task is a natural 'set' of items prompting the student to conduct a scientific investigation associated with a range of NSES inquiry standards and linked to a common stimulus.

Choosing a Student Model

Our goal of developing a Task Specification for Floating Pencil required that a Student Model be defined. Specifically, the Task Specification requires a summary of the Student Model as well as definitions of Student Model Variables.

Initially, our group explored different aspects of the Student Model. Previous research indicated that Floating Pencil is content lean and inquiry constrained (Bass, Magone, & Glaser, 2002; Baxter & Glaser, 1998). We determined that much of the needed content knowledge is given in the task directive. It was apparent that the inquiry process for Floating Pencil was highly scaffolded; specific directions were given for each step of the investigation. Therefore, we concluded that the task offered evidence for students' inquiry skills – distinguishing correctly implementing investigative procedures from incorrectly implementing procedures. Team members examined the cognitive demands of each item in Floating Pencil; we discussed whether a set of instructionally-linked Student Model Variables (SMVs) could be created, based on each item.

Different ways of defining the Student Model for Floating Pencil were explored. We noted that in the PADI Design System, multiple Student Models can be identified in Templates, whereas Task Specifications allow only one Student Model. The types of Student Models that were considered were: (1) the NAEP framework,[3] (2) instructionally-based variables (a multi-dimensional model that we would develop ourselves), (3) NSES inquiry standards (a multidimensional model based on the phases of inquiry assessed by Floating Pencil), (4) a 2-dimensional content by inquiry model, and (5) a 1-dimensional model measuring science proficiency.

Our team connected the Student Model as closely as possible with NAEP's purposes and framework. This allowed every Floating Pencil Activity to be linked to an already existing Student Model Variable and was consonant with the practices of the other PADI strands. Both BioKIDS and FOSS matched each of their items with the developer's framework within the Student Model. The content and process codes from NAEP's framework became the underlying variables of our Student Model. As we proceeded, we tried to determine NAEP's assessment argument for Floating Pencil and NAEP's definition of inquiry. We found it difficult to clarify the cognitive model that underlies the Floating Pencil task.

As the Task Specification was drafted, the Student Model was refined. Our team considered dropping 'conceptual understanding' as a process dimension, because every item is coded for one of three content areas. In doing this, there was some discussion as to whether conceptual understanding was distinct from types of content knowledge. Secondly, we dropped the content area 'life sciences' from our Student Model because none of the Floating Pencil items, nor the larger group of items[4] Floating Pencil was to be calibrated with, were life sciences items. It was noted that dropping this content area gave us only a subset of the NAEP

[3] The NAEP framework is a 3 X 3 content by process matrix. The content categories are physical science, earth and space science, and life science; the process categories are conceptual understanding, practical reasoning, and investigation. Every NAEP science item is given one content code and one process code.

[4] The Floating Pencil task will be calibrated with other items from the '*Validities of Standards-Based Science Inquiry Assessments: Implementation Study*'

framework. Finally, we made a substantial refinement to our Student Model Variables. Six SMVs were defined based on every possible content / inquiry pairing in the Student Model:

- physical science conceptual understanding
- physical science practice reasoning
- physical science investigation
- earth and space science conceptual understanding
- earth and space science practice reasoning
- earth and space science investigation

The Evidence Model

This section will address the development of both the Evaluative Sub-Model and the Statistical Sub-Model of the Evidence Model. Please note that what is known as the Statistical Sub-Model in the ECD framework is referred to as the Measurement Model on the PADI system; both of these terms refer to psychometric models.

Determining Floating Pencil Structure

The Floating Pencil task is comprised of 14 items and is accompanied by a NAEP Scoring Rubric. Based on the Rubric, some sets of items are scored together; items 3, 4, 8, and 11 are scored together, and items 5, 8, and 11 are scored together. Characteristics of the Floating Pencil Task were explored through the lens of PADI representational forms (Template, Design Pattern, Task Specification). Initially, our group explored the features of the Floating Pencil items by specifying Materials and Presentations (M&P), Work Products, and Observable Variables (OVs) in the draft Template. For the purposes of drafting our Template and Task

Specification, we considered items scored together as one Activity. More specifically, any item or group of items resulting in a single Observable Variable would be represented within a single Activity. For Floating Pencil, we defined 10 Activities on the basis of the 14 items. Aspects of the common stimulus, the directive and the physical stimulus materials, also were explored. The directive, we determined, belonged in the category of Materials and Presentation. The physical stimulus included three bottles of solutions, a graduated cylinder, a pencil with a thumbtack in the eraser, and a ruler in centimeters. We noted that the common physical stimulus materials created item dependencies. Sequential dependencies were another type of dependency found in the Floating Pencil task; these are situations in which the response to one or more items serves as part of the stimulus for a subsequent item. Work Products were identified for Floating Pencil; these are a multiple-choice response, essay responses, numerical responses, numerical entries in a table, plotting points on a graph, and drawing a line on a graph. Observable Variables are scored Work Products, based on the Evaluative Sub-Model. In addition to the scores that NAEP identified, we also considered other possible scores resulting from the Floating Pencil task, including an overall 'science proficiency' score.

A critical decision in structuring the Task Specification had to do with the relationship between Templates and Activities within the Floating Pencil task. Our group discussed whether to have 1 template with 14 Activities, 1 per item; 14 Templates, 1 per item; or 1 Template with 10 Activities defined on the basis of the Observable Variables. We decided to pursue 1 template with 10 Activities. The reason for this was the conditional dependency of all the Floating Pencil items on a common task directive and physical stimulus.

Developing an Evidence Model for the Task Specification

The Evaluative Sub-Model for Floating Pencil was based on the NAEP Rubric. Two group members specified our Evaluation Procedures in the Task Specification. The mapping between Work Products and Observable Variables was based on the NAEP Rubric. In doing this, we identified and added to the PADI Design System three new Work Products, Numerical Response, Table Entry – Numerical Response, and Graphical Elements.

In terms of the psychometric model our group came to understand that using the BEAR Scoring Engine (as other PADI strands have done) to score our Floating Pencil data would limit the analysis to the MRCML model, a multidimensional Rasch model (Kennedy, in press). Prior to our understanding that the BEAR Scoring Engine supported only an MRCML model, we considered other potential psychometric models for the data. For example, we considered a 3PL model to estimate slope and guessing parameters. In the future, other scoring engines are likely to be developed. This was clarified on the PADI Design System by updating a section of the glossary. One implication of this limitation is that guessing and item discrimination parameters cannot be estimated currently with the PADI Design System.

In the PADI Design System, the Measurement Model is defined as part of the Template and Task Specification. As with the Student Model, our group sought to closely mirror NAEP's approach. Our Student Model Variables were based on the NAEP framework. In terms of scoring, we tried to understand the meaning of NAEP's reported scores and found that the proportion of students at each score level were reported for the observable variables within Floating Pencil. The items in Floating Pencil did not appear to be included in NAEP's IRT analyses Thus, we were unable to determine whether NAEP found the scores to be statistically dependent or independent. The psychometric qualities of the Floating Pencil items will be examined in the subsequent data analysis phase of our work. We sought to clarify some elements of the Measurement Models as prompted by the Task Specification. The purpose and values of Mins and Maxs for the SMV scales were clarified, as were the scales for SMVs (continuous) and OVs (categorical). The Student Model for the Floating Pencil Task Specification is a multidimensional Measurement Model,though each of the 10 Activities do not employ all the dimensions. Thus, we had to learn how to code those SMVs that were not present in the Measurement Model for each Activity.

After the Student Model was refined (combining content and process dimensions), we considered some implications on the Evidence Model. It was noted that only a single dimension or Student Model Variable enters the Measurement Model for a single activity. Thus, there would be no "within item" multidimensionality to model in future analyses. Since different activities tap different SMV's, the Measurement Model would need to account for the "between-item" multidimensionality.

Finally, our team considered some future directions for measurement work. This included plans to work with alternate Student and Measurement Models (e.g., a 3PL model). An important direction that was raised is taking conditional dependencies into account, perhaps through an item bundling study (Wilson & Adams, 1995). Lastly, we may give some consideration to psychometrically 'parallel' tasks.

The Task Model

Selecting a Set of Assessment Tasks

The set of assessment tasks our team selected to be reverse engineered into a Task Specification would have to have a number of characteristics. First, the task or set of items would be selected from a pool of NAEP and other released science items. Second, we would have to select a set of assessment tasks that were useful for measuring inquiry. Third, these tasks would have to be connected as themes, blocks, performance assessments, and/or based on some set of common qualities. Fourth, these tasks could be of the type multiple-choice, hands-on performance assessment, open-ended, or scenario-based, or a mix of types. Different mixes of items would have different surface and deep characteristics. The Task Model and the Task Model Variables (TMVs) are where the surface and deep characteristics of items are represented. Examples of Task Model Variables include the number of distractors in multiple choice items, scaffolding levels, number of stimuli present, and use of graphics. In defining TMVs, we would consider the links among the characteristics of the tasks and the Observable Variables.

Developing a Task Model

Initially, group members examined Floating Pencil assessment tasks by taking the performance assessment and sharing their experience and thoughts about the assessment tasks with team members. As a whole, the task was found to be scaffolded procedurally, but not in terms of content, and the reading demands were high. We noted that the Materials and Presentations used in the assessment task were likely to impact the difficulty of the task. One example is that the pencil representation in the test booklet was not drawn to scale, making the students' choices of using the representation or the actual pencil itself highly consequential. Another example is the ruler that is given to students. If a ruler is used that has several scales on it, the task would be made more difficult.

Our group explored the characteristics of the Floating Pencil Task through the lens of PADI representational forms (Template, Design Pattern, Task Specification), which requires that we define TMVs. We discussed how variations in the number of solutions, a thicker pencil, and different water temperatures would impact the assessment task. We discussed the importance of standardizing the materials for Floating Pencil to eliminate sources of error and the impact of using physical materials as a common stimulus. We discussed the research results that identified the task as content-lean and inquiry constrained (Bass, Magone, & Glaser, 2002; Baxter & Glaser, 1998). This led us to distinguish tasks as content lean or rich, and the process as closed or open. In completing the Task Specification, we identified the following key TMVs: the physical materials of the task, the level of inquiry structure (including scaffolding, difficulty, and extensiveness of inquiry), the level of content structure (also including scaffolding, difficulty, and extensiveness), the verbal demand of the task, and the cognitive complexity of the task. In working with the Task Specification, our group questioned which TMVs are specific to an activity and which are set at the template level.

Among the issues we considered were whether the task directive should be considered a TMV or part of Materials and Presentations, how TMVs can be sources of measurement error, and the influence of features of the stimulus materials on Observable Variables. Our team discussed how issues of the context of the assessment (e.g., student prior knowledge of content topics) are handled by a PADI Wizard, not the Task Specification. The relationship between Student Model and Task Model was considered; it was noted that the Student Model puts boundaries on what can vary in different instances of a task. In addition, we came to understand that there are different kinds of TMVs. TMVs can play many roles (as many as 10, it was suggested), and a single TMV can play multiple roles. TMVs may or may not be connected to the Student Model; TMVs that are not connected to the Student Model can serve as alternative hypotheses and are potential sources of measurement error. Finally, we gave consideration to the process of narrowing down our list of TMVs.

There was an ongoing interplay between our consideration of Task Model Variables, conception of the 'family of tasks' of which Floating Pencil was a member, and thinking about additional Templates and Design Patterns for inquiry assessments that would be of a broader scope than a Task Specification. At the core of identifying TMVs and conceptualizing broader PADI representations was the Floating Pencil question, "this task is an instance of a what?" (or alternatively, "what is a parallel version of this task?"). This question stimulated multiple conversations about the family of tasks of which Floating Pencil is a member. It was agreed that such a family of tasks was not to be discovered, but constructed. Towards this end, we analyzed a 4th grade version of the Floating Pencil task. In comparison with the 8th grade version of the task, the physical materials were similar, the directives differed, the number of activities differed, the number of skills assessed differed, and the cognitive complexity of the activities differed. Given the assumption that the 4th grade and 8th grade versions belonged in the same family of tasks, we postulated some additional TMVs for the Floating Pencil task: the numbers of activities, the numbers of skills assessed, the sequencing of cognitive complexity across the task, the presence or absence of graphing, the number of measurements taken, and the number of salt solutions. Subsequently, a family of tasks was defined based on already existing NAEP-like performance assessments. It was argued that a family of tasks should be based on a Student Model and assessment argument. Towards this end, we will consider creating one or more generalized templates based on phases of inquiry.

Discussion

As part of the PADI project, our Floating Pencil team was involved with an extensive networked collaboration involving a Web-based work place tool. Common ground and shared understandings of our work were established through regular meetings (weekly meetings with the entire PADI team and regular meetings of our group) and through interactions with the PADI Design System and support materials. During the meetings, group members negotiated their understandings of the Floating Pencil task in interaction with the PADI Design System and underlying ECD framework. In addition, team members, both jointly and individually, accessed resources from the PADI Design System including exemplar representations (e.g., previously authored Templates), system tools, and other resources (e.g., PADI technical reports).

The PADI Design System also served as a repository of emerging knowledge; for example, team members drafted portions of the Task Specification with ample comments, and other team members responded, on-line. The coupling of work, involving the relationships between individual and joint efforts, was handled by Geneva Haertel, one of the Co-PIs. Geneva balanced the affordances of team members' expertise, group and individual learning curves with the PADI Design System, and the developmental needs of the project itself. Technological readiness was assured by the nature of the PADI Design System. The Design System, including our draft representations, was immediately available to all group members. This allowed everyone to view the same information on-line; for example, changes to the Task Specification were immediately available to all members who accessed the Internet.

The results of our work on the Floating Pencil task include the analysis of the task itself, the creation of a 'trace' of our work in the form of a Task Specification, the impact of our work on the PADI Design System itself, and some understanding of the design space of NAEP test developers. Our analysis of Floating Pencil was largely an epistemological process in which our team came to understand more deeply the features of the Floating Pencil task, the ECD framework, and the PADI Design System (as well as the interactions among these areas). Our work left a ‘trace’ in the form of the Floating Pencil Task Specification. This PADI representation is now available as part of the PADI library; this is the first blueprint based on a large-scale science assessment task. The Floating Pencil Team’s work also impacted the PADI Design System. For example, a number of new Work Products were added to the Design System to accommodate the Floating Pencil task. Another example is that our team’s work prompted discussions about redundancies on the PADI system – redundancies at the level of Design Patterns, Work Products, and Task Model Variables. Steps were taken in the PADI project as a whole to address these redundancies. Finally, we learned some things about the design space of NAEP Test Developers. NAEP Test Developers were given a ‘loosely coupled’ framework with which to describe students’ knowledge, skills, and abilities (each item is to belong to one and only one content category, and one and only one process category). The level of specification for science content is very large-grained; we do not yet know how this will play out with our measurement model.

The Floating Pencil group will continue its work over the course of the next year. We will complete the definition of a task family for Floating Pencil and clarify Task Model Variables at the Activity and Task levels; this may be done in the form of developing abstract templates that span the phases of scientific inquiry. Our group will also conduct empirical data analyses using the MRCML multi-dimensional model and NAEP framework. A technical report of our work on Floating Pencil is underway. Finally, we will conduct additional data analyses using other Student Models.

References

Brown, A. L., Ash, D., Rutherford, M., Nakagawa, K., Gordon, A., & Campoine, J. C. (1993). Distributed expertise in the classroom. In G. Salomon (Ed.), *Distributed cognition*. New York: Cambridge University Press.

Bass, K. M., Magone, M. E., & Glaser, R. (2002). *Informing the design of performance assessments using a content-process analysis of two NAEP science tasks* (CSE Technical Report 564). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Issues and Practice*, *17*(*1*), 37-45.

Gee, J. P. (2000). Communities of practice in the new capitalism. *The Journal of Learning Sciences*, *9*(*4*), 515-523.

Greeno, J. G. (1999). Situative research relevant to standards for school mathematics. Prepared for a volume of papers edited by J. Kilpatrick, *Reviewing research relevant to the revision of the NCTM Standards*. Stanford, CA.

Greeno, J. G., Collins, A. M., & Resnick, L.B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York: Macmillan.

Greeno, J. G. & T. M. Group (1997). Participation as fundamental in learning mathematics. In J. A. Dorsey, J. O. Swafford, M. Parmantic, & A. E. Dossey (Eds.), *Psychology of mathematics education* (Chap. 1). Columbus, OH: Eric Clearinghouse for Science, Mathematics, and Environmental Education.

Kennedy, C. (in press). Constructing PADI measurement models for the BEAR scoring engine (PADI Technical Report 7). Menlo Park, CA: SRI International.

Luff, P, Heath, C., Kuzuoka, H., Hindmarsh, J., Yamazaki, K., & Oyama, S. (2003). Fractured ecologies: Creating environments for collaboration. *Human-Computer Interaction*, *18*, 51-84.

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

Mislevy, R. J., Hamel, L., Fried, R., Gaffney, T., Haertel, G., Hafter, A., et al. (2003). *Design patterns for assessing scientific inquiry* (PADI Technical Report 1). 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., Almond, R. G., Haertel, G. D., & Penuel, W. (2003). *Leverage points for improving educational assessment* (PADI Technical Report 2). Menlo Park, CA: SRI International.

Olson, G.M. & Olson, J.S. (2000). Distance matters. *Human-Computer Interaction*, *15*(*2&3*), 139-178.

Quellmalz, E. S., Haertel, G. D., Lash, A. L., et al. (2004, April). *Evaluating students' opportunities to learn the science content and inquiry skills measured in large scale assessments*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Quellmalz, E., & Haydel, A. M. (2003, April). *Using cognitive analyses to describe students' science inquiry and motivation to learn*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Quellmalz, E., & Kreikemeier, P. (2002). *Validities of science inquiry assessments: A study of the alignment of items and tasks drawn from science reference exams with the National Science Education Standards*. Paper presented at the annual meeting of the American Education Research Association, New Orleans.

Schank, P. & Hamel, L. (2004) *Collaborative modeling: Hiding ULM and promoting data examples in EMo*. Presented at the Computer-Supported Collaborative Work Conference, Chicago. November, 2004.