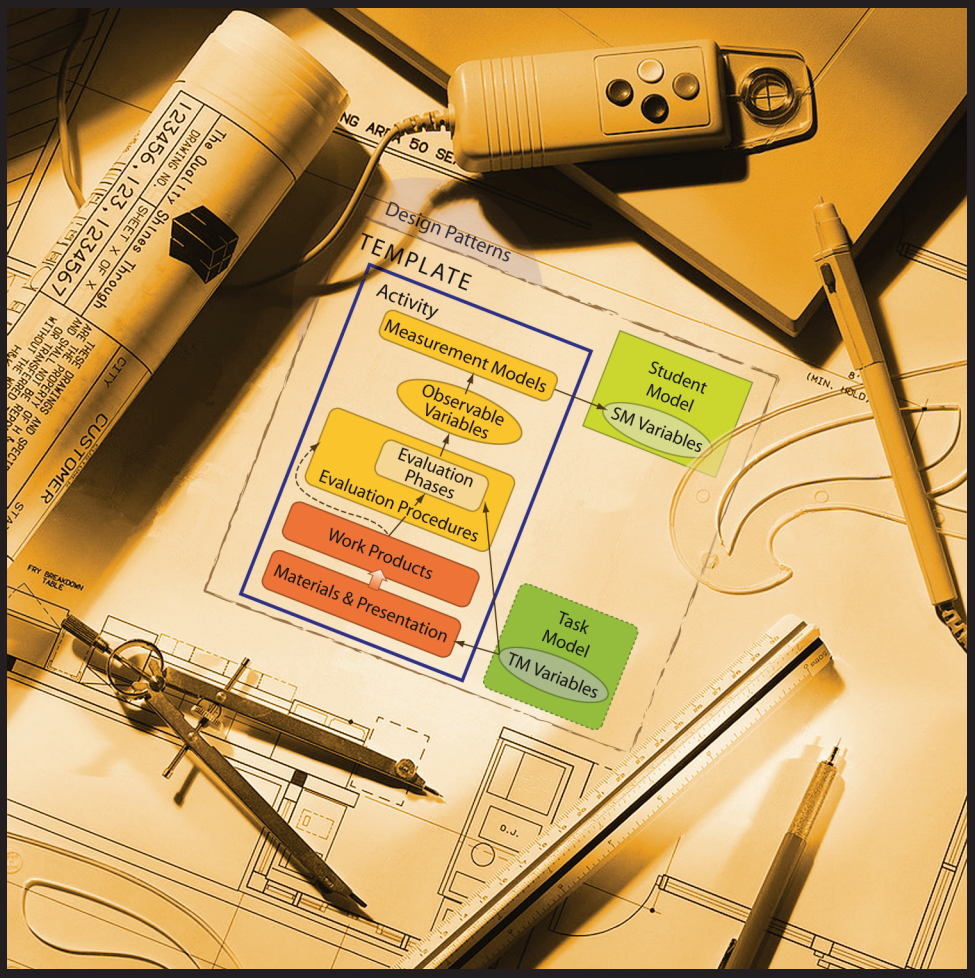


Large-Scale Assessment Technical Report 5| February 2010

Planning Evaluations for DR K-12 Design Projects: An Evaluative Framework Using Evidence-Centered Design

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

Kathleen Haynie, Haynie Research and Evaluation

Report Series Published by SRI International

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

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

APPLICATION OF EVIDENCE-CENTERED DESIGN TO STATE LARGE-SCALE SCIENCE ASSESSMENT
TECHNICAL REPORT 5

Planning Evaluations for DR K-12 Design Projects: An Evaluative Framework Using Evidence-Centered Design

February 2010

Prepared by:
Kathleen Haynie
Haynie Research and Evaluation

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

Contents

Tables

Figures

Abstract

This technical report provides the evaluation design for the DR K-12 project, “Application of Evidence-Centered Design for State’s Large-Scale Science Assessment”. Evaluation work is presented and discussed in terms of evaluative requirements from the National Science Foundation, and in terms of five fundamental issues that undergird practical program evaluation: social programming, knowledge construction, valuing, knowledge use, and evaluation practice. The reports moves through six phases of evaluative work: (1) a logic modeling process, (2) definition of evaluative focus areas, (3) clarification of evaluative questions, (4) design of the evaluation plan, (5) collection and analysis of data, and (6) provision of evaluative information to stakeholders. Rationales for design decisions made in the context of this evaluation are provided and discussed, and process options are laid out for designers of similar evaluations.

1.0 Introduction

This technical report will describe the process of designing the evaluation of the DR K-12 project, Evaluation of Application of Evidence-Centered-Design to State's Large-Scale Science Assessment. I hope that this report will serve to clearly explicate the evaluative design process for this project – pointing the way for the evaluation of other subsequent DR K-12 projects, particularly those using evidence-centered design. Throughout this report, the reader will come to understand the particular design choices made, resulting in an array of methodological approaches to this evaluation. The intention is for this report to be informative to evaluators, educational researchers, psychometricians, and test designers. Through engaging with this report, the reader can come to understand more deeply the evaluative process for this project, as well as to think more deeply about an overall evaluative approach and potential evaluative design choices for similar projects.

1.1 Framing of Evaluation

Evaluations of social programs have been conducted for the last five decades with the purposes of program improvement, and hence, the improvement of the welfare of individuals, organizations, and society. Evaluative efforts aim to provide systematic feedback on programs in terms of agreed-upon criteria. Thus, evaluations have particular characeristics and qualities, much like any type of research endeavor. Weiss (1998) defines evaluation as "the systematic assessment of the operations and/or outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy" (p.4). But the devil is in the details -- what the criteria are, who establishes them, how the evaluation is designed, what counts as evidence, and how evaluative feedback is communicated to stakeholders and utilized. Evaluations are not carried out in a vacuum –- they are part of a complex, interdependent, nonlinear set of problem-solving activities –- and, as such, are situated within the affordances and constraints of an emerging or existing program.

Practical program evaluations are needed to shape emerging projects, such as a DR K-12 project, and such evaluations have particular characertistics. Five fundamental issues undergird practical program evaluation, according to Shadish, Cook, and Leviton (1991). These are social programming, knowledge construction, valuing, knowledge use, and evaluation practice. Social programming is the ways that social programs and policies develop, improve, and change, especially with regard to social problems. Knowledge construction is the ways researchers or evaluators learn about social action. Valuing is the ways that value can be attached to program descriptions. The ways social science information is used to modify programs and policies is knowledge use. Finally, evaluation practice is defined as the tactics and strategies evaluators follow in their professional work, especially given the constraints they face. In the text that has come to serve as a foundation for the field of evaluation, Shadish, Cook, and Leviton (1991) analyze existing theories of evaluative practice in terms of these five fundamental issues. Extending this

focus, a given program evaluation or evaluation design might be analyzed according to these dimensions. This broad lens will inform our consideration of the DR K-12 project evaluation design.

1.2 Introduction to DR K-12 Project

In August 2006, the National Science Foundation posted solicitations for a new program called Discovery Research K-12 (or DR K-12). This DR K-12 program was created to support research, development, and evaluation of knowledge generation and application to improve K-12 learning; the DR K-12 solicitation represents a consolidation and re-alignment of the Teacher Professional Continuum (TPC), Instructional Materials Development (IMD) and Centers for Learning and Teaching (CLT) programs that were administered in the Division of Elementary, Secondary, and Informal Education. In the first year (FY2007), a total of $42,000,000 in funding was available for about 48 projects: 12 Conference, 21 Exploratory, and 15 Full-Scale. Activities were funded in three major areas:

- *Applied Research* that supports three categories of projects: Evaluative Studies of NSF-Funded Resources and Tools, Studies of Student Learning Progressions, and Studies of Teachers and Teaching.
- *Development of Resources and Tools* that supports two categories of projects: Assessment of Students' and Teachers' Learning and Instruction of K-12 Students and Teachers.
- *Capacity Building* that supports two categories of projects: STEM Systems Research and STEM Education Research Scholars.

Subsequent DR K-12 solicitations have been offered by the NSF in FY2008, FY2009, and FY2010.

This DR K-12 project, Application of Evidence-Centered-Design to State's Large-Scale Science Assessment, was funded as a five-year, full-scale project in FY2007, under the specialty area Development of Resources and Tools – Assessment of Students' and Teachers' Learning. A goal of the present project was to illustrate the use of evidence-centered design in the Minnesota Science Assessment (MCA-II) in ways that not only benefit the MCA-II, but provide value for the larger science education and assessment communities. As a stipulation for project funding, an external evaluation was required, involving an external evaluator and an expert advisory panel. The NSF required this evaluation to begin with the commencement of the project – allowing evaluation design to occur concurrently with project planning and activities. This technical report will describe the evaluative planning process and underlying rationale in the first two years of the project (September 2007 through September 2009).

2.0 Evaluation Requirements And Process

2.1 NSF Requirements for Evaluation

In the request for DR K-12 proposals due January 2007 (National Science Foundation, 2006), NSF outlined evaluative requirements for DR K-12 projects initiated in 2007. NSF's explicit statement of these requirements reflects its evaluation policy, though this policy might not be considered as formal (Mark, Cooksy, & Trochim, 2009) since no mechanisms are in place for its enforcement. NSF stated:

> *All projects are expected to include an evaluation plan that examines the extent to which the project has met its goals. The proposal should describe how the objectivity of the evaluation will be ensured. Summative components of evaluations must be conducted by a researcher or evaluator external to the project and submitted with the NSF final project report. The proposal should specify the evaluation questions, the methods to be used, the data to be gathered, and the data analysis plans. Responsibilities should be clearly defined. For formative evaluation, plans should address how appropriate feedback will be given to the project leadership team so that it can make modifications to the project activities and address significant issues in the annual report…There will be a third-party DR K-12 program evaluation designed and implemented by external evaluator(s) to track the program's progress in meeting overall goals. All projects are expected to collaborate with this program evaluation.*

Additionally,

- *All projects must have an evaluation plan, including performance indicators and other specific measures that will be used by the project team to assess the project's progress and success in meeting its goals and objectives.*
- *The evaluation plan should address but not be limited to the following methodological considerations:*
    - *Poses significant questions that can be addressed empirically and that are central to the project's goals and objectives as well as contributing to understanding that meets current and expected educational demands of the nation on world-class criteria.*
    - *Reflects clearly how current literature and the context of the project inform the evaluation methodology and goals.*
    - *Plans for evaluation and/or research methods appropriate to the questions posed and to possibly emergent questions that arise during the course of the project with a credible rationale for selection of methods.*
    - *Provides clear and logical arguments and evidence for conclusions drawn and addresses plausible rival interpretations of findings.*

- *Makes use of existing data where possible and takes into account ways of reducing the burden on people and institutions in data gathering.*
- *Contributes to understanding of what factors contribute to the project's success in meeting its goals and objectives and understanding of the effects of the evaluation itself*

And finally, *Resources and Tools* projects are required to have an evaluation plan that covers all critical components of the project, including formative assessment of the development process (which may be conducted by an internal evaluator) and summative evaluation (which must be conducted by an external evaluator) that speaks to factors affecting implementation as well as data and analysis from pilot- and field-test results. In addition, as part of the evaluation, all materials must undergo independent review by qualified experts to ensure accuracy of the content, appropriateness of the pedagogy, and suitability of the contexts, language, etc., for the intended audience.

NSF's evaluation requirements clearly provide a set of important considerations for evaluation design. In particular, the expectations include:

- Developing an evaluation plan
    - designed to examine the extent to which the project has met its goals, including performance indicators and other measures
    - reflective of the current literature and project context
    - addressing the development of materials and factors affecting implementation
    - that specifies evaluation questions, methods (appropriate to questions), data (existing, where possible), data analysis plans, and logical arguments for evidence-based conclusions
    - that specifies use of formative evaluation and how formative feedback will be communicated with the project team
    - that contributes to understanding what factors contribute to project success
- Use of a third party (external) evaluator and insurance of objectivity
- Independent review of all materials for quality and appropriateness with respect to the intended audience

Evaluation requirements from the NSF are clearly related to the core dimensions laid out in the introduction to this. Let's begin with the requirement of an evaluation plan. A critical element of evaluative practice involves the narrowing of options to do a feasible evaluation, given time and resource constraints. An evaluation plan can clearly and explicitly lay out the scope of the evaluative work (i.e., what will and will not be included), thus reflecting decisions of where to focus the evaluation among many competing options. The NSF also asks that the evaluation plan state how the extent to which the project

has met its goals will be studied - a requirement indicating project goal fulfillment as a core value. Thus, the funder has specified a particular summative outcome, goal fulfillment (however, goal definition is left up to the project, pending NSF's acceptance of the proposal), although the valuing (the third core dimension previously laid out)) of particular goals comes from the proposal development team (and the project team, if the goals are subsequently revised). The valuing of the goals as outcomes (instead of valuing different outcomes or placing more emphasis on process evaluation), as prescribed by NSF, constrains the evaluation design. In addition to goals, evaluation of the implementation of the development process for resources and tools is a critical requirement, again reflecting NSF's values for DR K-12 projects. NSF requires the evaluation plan to reflect the current literature and project context – thus placing the project in the larger context of educational practice and advances. Choices of aspects of the evaluation to be included in the plan is an aspect of evaluative practice; specifying these aspects also bears on the knowledge component of evaluation – the assumption that knowledge will be more reliably constructed if questions, methods, and data analysis approaches are specified in the evaluation plan. NSF also asks that the evaluation plan specify significant evaluation questions that can be addressed empirically (an aspect of knowledge construction as well as evaluative practice). The specification of methods appropriate to the questions posed is also an aspect of knowledge construction – seeking methods that provide valid evidence informing critical questions. The requirement of using existing data, where possible, is an aspect of evaluative practice that bears on efficient use of resources and respecting collaborative relationships within project-based work. Providing clear and logical arguments (and addressing rival hypotheses) again bears on knowledge construction and the validity of knowledge claims. NSF also articulates requirements for responsibilities and for offering formative feedback. Their statement that assignments and responsibilities should be clearly defined, an aspect of evaluative practice, assumes that clear specifications will lead to better evaluative practices. The use of evaluative results by the project team, and facilitation of this use by the external evaluator, is considered to be of great importance by the NSF; NSF asks that formative evaluation plans address how appropriate feedback will be given to the project leadership team so that modifications to the project activities can be made and significant issues can be addressed in the annual report.

The NSF states that the proposal should describe how the objectivity of the evaluation will be ensured, which bears on knowledge construction and evaluative practice. Ensuring the objectivity of the evaluation is a question of methodology and evaluative practice. Methodologically, the methods that are used to gather evidence that inform evaluative questions, the claims that are made, and the warrants for the claims that are made all impact the objectivity of the evaluation. However, objectivity is also a function of evaluative practice – ensuring that the evaluator collects, analyzes, and communicates information in an objective way. NSF further requires that summative components of evaluations be conducted by a

researcher or evaluator external to the project and be submitted with the NSF final project report. This stipulation, an aspect of evaluative practice, directs the evaluation to:

- have an adequate focus on summative evaluation
- employ an external evaluator or researcher (who is less likely to serve the interests of the funded organization)
- include "non-summative" components, by direct inference
- prepare summative evaluation work for the final NSF project report

2.2 Evaluation Overview

The evaluation process for this DR K-12 project was designed by the external evaluator, Dr. Kathleen Haynie (Director of Haynie Research and Evaluation) to meet the requirements of the NSF (see previous section), as well as the needs of this particular project. Dr. Haynie's expertise is in the areas of educational psychology, assessment, science education, and evaluation. As stated in the proposal, evaluation plans have been reviewed by an expert Advisory Panel[1]. In accordance with best practices in the evaluation field, NSF requirements, and project needs, the evaluation is designed to unfold in six phases:

(1) a logic modeling process,

(2) definition of evaluative focus areas,

(3) clarification of evaluative questions,

(4) design of the evaluation plan,

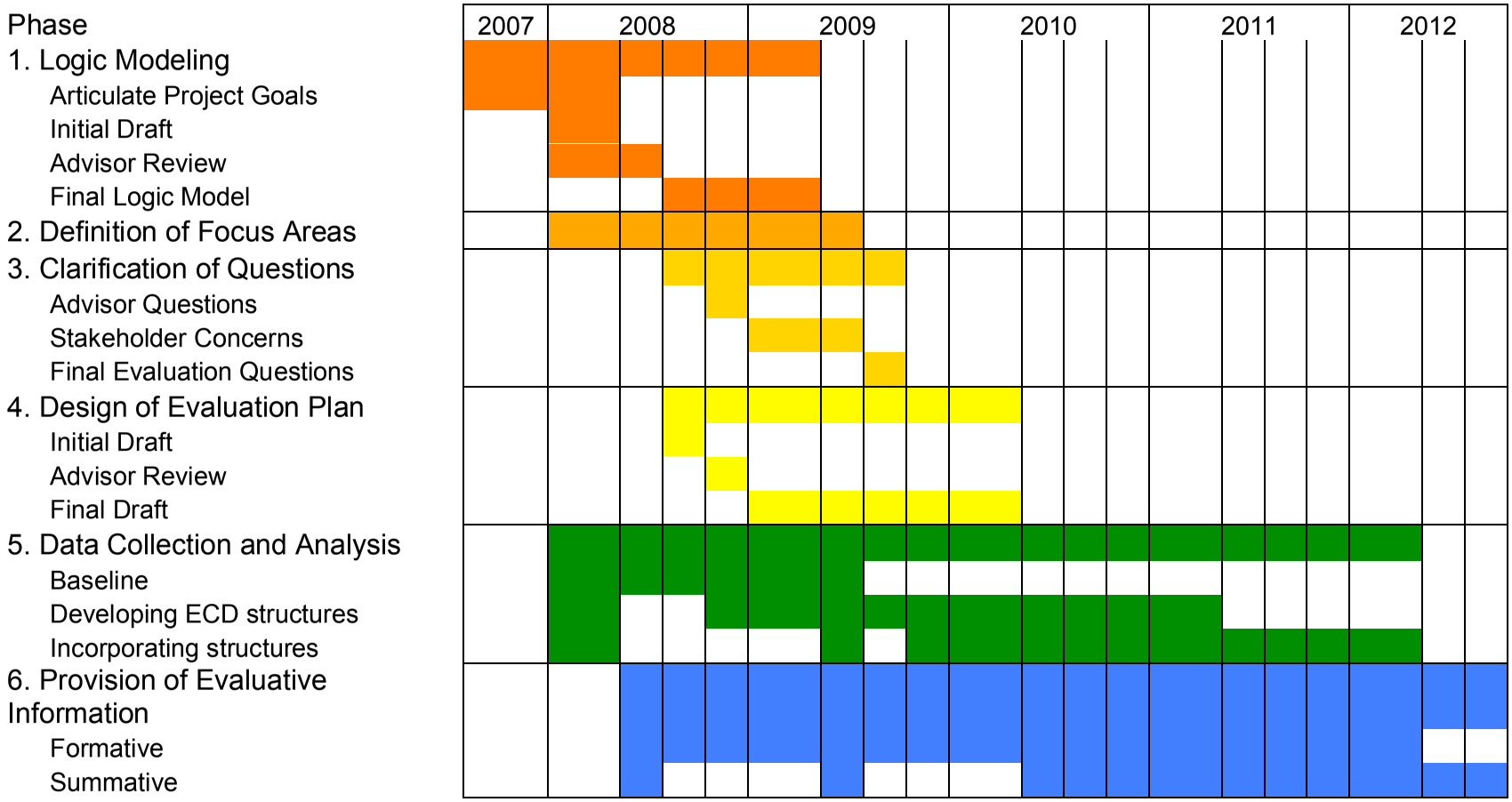
(5) collection and analysis of data, and

(6) provision of evaluative information to stakeholders.

The expected timeline for the evaluation phases, which are discrete but often concurrent, is provided in Figure 1. This timeline spans the length of the project – from September 2007 through August 2012. We will now discuss each phase of the evaluation.

[1] Members of the Advisory Panel are Dr. Jamal Abedi (UC Davis), Dr. Greg Chung (UCLA), Dr. Edward Haertel (Stanford University), Dr. Kristen Huff (College Board), Dr. Cathleen Kennedy (BEAR Assessment Center, University of California, Berkeley), Dr. Steve Robinson (TN Tech), Dr. Nancy B. Songer (University of Michigan), and Dr. Chris Swoyer (University of Oklahoma). This panel was chosen to provide overall guidance for the summative evaluation and review all performance indicators and evaluation instruments. The combined expertise of this Panel spans the areas of psychometrics, ECD, educational psychology, science teaching and learning, philosophy of science, and previous NSF-funded work (i.e., the PADI project).

Figure 1. Timeline for the Evaluation



3.0 PHASES OF THE EVALUATION

3.1 Phase 1 – Logic Modeling

A project logic model and associated theory of change were developed during the first year of this project (2007-2008) to guide program planning, administration, evaluation, and research goals. At the outset of this phase of work, team members articulated the goals of this study (see Table 1). During this time, the external evaluator facilitated team members' refinement and elaboration of the logic model, explication of program theory (i.e., why the project will be able to accomplish its goals) and development of shared understandings of the relationships between project activities and outcomes. As part of the evaluation effort in the first year, the project goals and logic model were provided to the Advisors and NSF Program Officer in January 2008. An Advisory Panel meeting (with the NSF Program Officer) was held in February 2008 for the purpose of reviewing and revising this theory of change. Subsequent revisions to the logic model were made in December 2008 by project team members. The resulting (and current) logic model is provided in Figure 2.

Table 1. Project Goals

1) Apply the evidence-centered design (ECD) conceptual framework/paradigm to the designing of a large-scale, state-level, high stakes accountability testing program (MCA-IIs) and test development cycles. Determine, via research and evaluative methodologies, if and how ECD-based ideas can be gainfully applied to a large-scale, state-level assessment. In doing this, the MCA-II processes will not be re-engineered; rather, leverage points will be identified and tried out as proof-of-concept to compare with present processes.
2) Develop and implement structures that are intended to improve efficiency in the test development process. This will be done in actual operational work for the MCA-II, but on a limited scale supported by the DRK-12 scope of work. In doing this, collect and analyze research evidence of changes to the process. This includes generative schemas for conceptual elements and reuse of operational elements and data structures. The motivations for the content of this work are the MCA-II assessment itself, and the broader science education community as reflected, for example, in the National Science Education Standards (NSES) documents.
3) Develop and implement structures that are intended to improve the validity of the test development process and products. This will be done in actual operational work for the MCA-II, but on a limited scale supported by the DRK-12 scope of work. In doing this, collect and analyze research evidence of changes to process quality (e.g., assessments arguments from Minnesota state standards and benchmarks more explicit and more scaffolded). In addition, collect evidence of test development products (e.g., storyboards, items, tests).
4) Extend design components, representations, and tools developed in the Principled Assessment Design for Inquiry (PADI) project[2] to support the efficiency and reusability of the assessment design process, and the articulation of assessment arguments, through a generative design layer that is articulated with the Pearson Educational Measurement (PEM) development infrastructure. These extensions will be demonstrated on a limited scale consistent with the project scope.
5) Develop the human capacity and understanding of the DRK-12 team and other test developers in terms of ECD-based ideas, the PADI Design System, and the Pearson/ Minnesota Department of Education (MDE) test development process. The focus here is on those staff members that are supported by the DRK-12 project.

[2] Information about the PADI project can be accessed at: http://padi.sri.com/

6) Disseminate research and evaluative findings to various communities (i.e., assessment, educational research, practitioner, state DOE, graduate student), particularly with respect to scaling up ECD-based ideas and PADI-based processes in state-level assessment contexts. The results of this project will have valuable implications for scaling up because our limited-scale implementations were carried out in a real operational context.

In developing the logic model, the project team came to articulate cyclical processes functions within different layers of test design (see Appendix A). The processes within this research cycle are: (1) developing human capacity, (2) identifying points for improved efficiency and quality, (3) developing or modifying structures and processes, (4) incorporating, on a limited scale, structures and processes with item and storyboard writers, and (5) analyzing and reflecting on the research cycle.

The logic modeling and goal articulation process, including independent review by the Advisory Panel, are important in contributing towards the NSF evaluation requirements. Firstly, without clear articulation of goals, the central thrust of the evaluation – examining the extent to which the project has met its goals – would be anchorless. In addition, the logic modeling process provides the project team and external evaluation with working hypotheses linking project activities to intended outcomes, thus contributing to initial understandings of relevant factors to project success. Lastly, attaining a mutually constructed and agreed-upon logic model, including project goals and outcomes, serves as a firm foundation for articulating significant evaluation questions.

Logic modeling, as part of the evaluative process, strongly connects with the five dimensions of evaluation outlined in the introduction to this report. Logic modeling contributes to social programming through serving as an articulation of the internal program structure – how the inputs, activities, and outputs to the project relate to each other. The logic modeling process, properly executed, builds consensus among project personnel, advisors, stakeholders on goals, outcomes, processes (activities) of project – thus reflecting agreement on the problem that is being solved and the method for solving it. This process lends itself to tracking project success in terms of meeting goals and objectives, limiting conditions, strategies for achieving goals, resources, and timelines – potentially serving as the core of a focused project management plan. Finally, logic modeling reflects group process and shared understandings – potentially increasing the buy-in of all stakeholders.

In terms of knowledge construction, the logic modeling process supports the definition of goals. In the case of this project, goals (Table 1) were defined broadly, leading to a potentially broad data collection effort. This is supported by evaluation theorists such as Cronbach who discuss the tradeoffs between bandwidth and fidelity (Shadish, W.R., Cook, T.D., & Leviton, L.C., 1991), recommending that priority be given to discovery and description early on in the life of a project. Finally, logic modeling begins to unpack the causal questions for this project; Cronbach states that “progress in causal knowledge consists partly

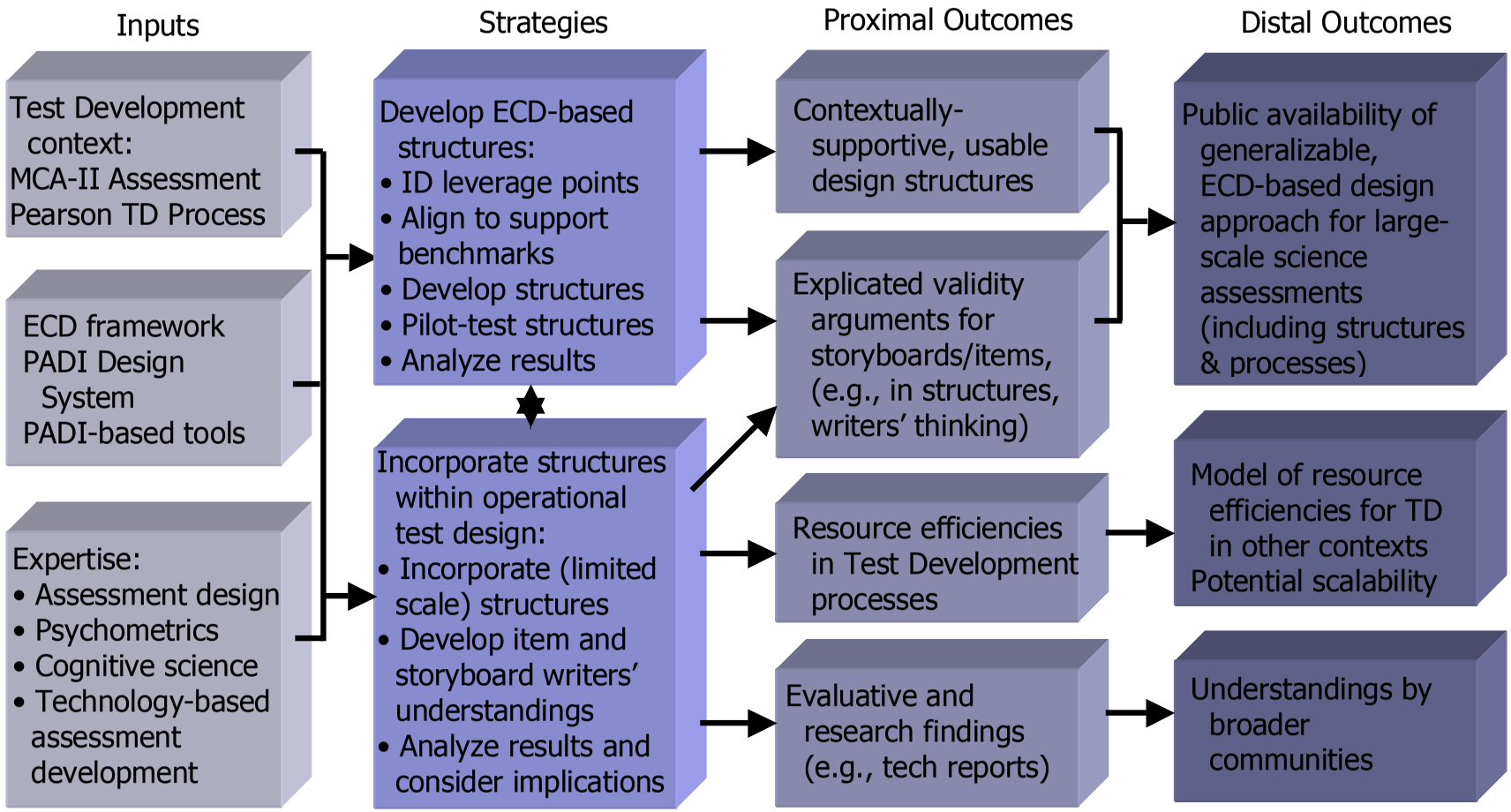
in arriving gradually at fuller formations" relationships among variables, the nature of manipulation, conditions surrounding it, characteristics of outcomes".

The articulation of goals is clearly valued by the NSF in conducting the evaluation; the project team itself chooses the content of those goals. In the logic modeling process, agreement and consensus are reached on what is valued as outcome and process. The evaluator can also choose their orientation to this goal definition process. As suggested by Weiss (1998), I choose to facilitate a process of goal construction from the values and goals that stakeholders (particularly the PIs), rather than allowing my own professional values to dominate.

The logic model can serve as the basis for the evaluation plan, involving stakeholders in logic modeling process so that buy-in and "ownership" occurs, which is central to the success of the subsequent evaluation effort and use of evaluative results. In leading the logic modeling process, I fostered collaborative development and discussion of logic model, frequent contact with users during (and after) the logic modeling process, and a focus on conceptual use of the logic model more than instrumental use.

Finally, in terms of evaluation practice, the logic model lays foundation for determining evaluative questions, methods, and decisions in practice. In the case of this project, a "discovery-oriented" project, the evaluation is (appropriately) not an experimental design, but prioritizes description and process. The design uses existing program data as a baseline for the "intervention" of the ECD-based tools. The evaluation, designed by the external evaluator, utilizes both a formative and summative approach, and the evaluator is responsible for facilitating the use of any evaluative results. This suggests a number of different roles for the external evaluator that might include designer, researcher, educator, diplomat, judge, reporter, designer, detective, and use advocate.

Figure 2. Project Logic Model



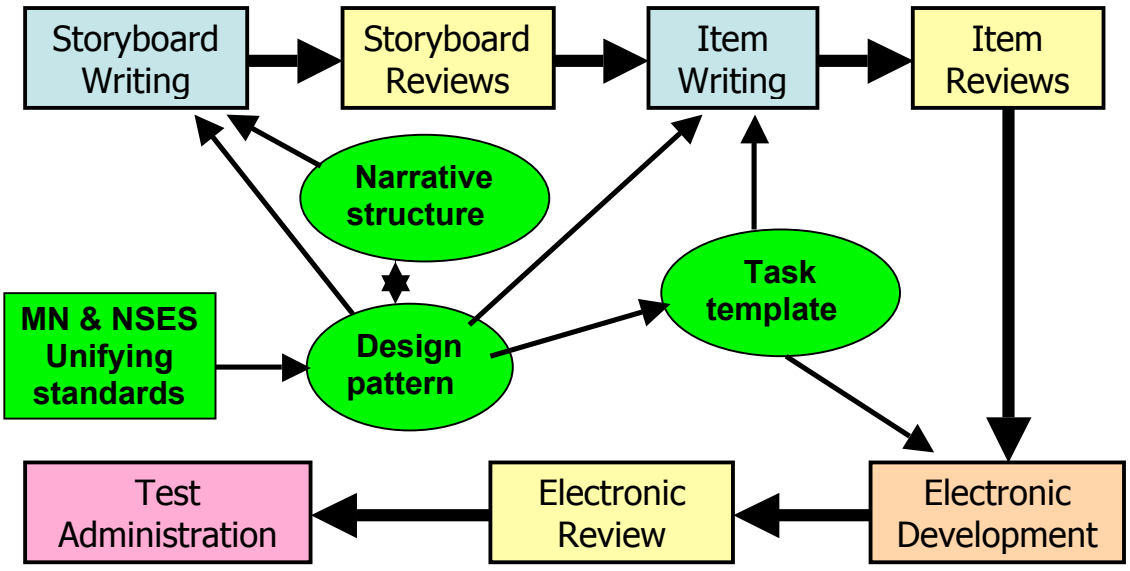
3.2 Phase 2 – Definition of Evaluative Focus Areas

During the first and second years of the project, I tentatively choose evaluative focus areas with respect to the logic model (see Figure 2), goals of the project (see Table 1), and project processes (see Appendix A). Evaluative focus areas were also based on the project team's articulation of "leverage points" for incorporating ECD-based structures and processes into the existing test development framework. These leverage points are depicted in Figure 3. The outer rectangular boxes represent the existing test development activity sequence, the green circles and boxes (in the center) represent new structures and information introduced by the project. The following leverage points can be seen:

- *design patterns*, informed by Minnesota and NSES standards, impacting storyboard and item writing, as well as task templates (narrative structures are an attribute of *design patterns*)
- *task templates* impacting item writing and electronic development

Key evaluative areas involve the development of ECD-based structures (e.g., *design patterns*, *task templates*), the review of the quality of resulting ECD-based structures, the piloting of these structures in non-operational contexts, the incorporation of these structures into operational assessment development, and the analysis of the quality of resulting products (e.g., storyboards, items).

Figure 3. ECD/PADI Leverage Points



Defining evaluative focus areas can serve as an antecedent to articulating evaluative questions (an important NSF requirement), and touches on important aspects of the evaluative process. Identifying the existing, operational test development process was critical for understanding the existing context into which this project must be aligned. This context, like any other, offers both constraints and affordances for the potential intervention (use of ECD-based structures), thus, helping to define in more detail the internal program structure as well as potential goals. Since causal knowledge is constructed on the basis of "fuller formations of relationships among variables, the nature of manipulation, conditions surrounding it…" (Shadish, Cook, & Leviton, 1991), knowing more fully the existing operational development process for this project, as well as identifying potential leverage points, moves us towards deeper understandings of potential causalities. Since realistic goals reflect both values and feasibilities, greater understandings of context may impact goals by fine-tuning them to the realities of particular contexts. In the case of this project, knowing the extent of operational storyboard and item development carried out in the given year made it clear that ECD-based structures could be developed to support some, but not all, areas of the MCA-II. Defining evaluative areas as a pre-cursor to developing evaluative questions served to be an important step in this evaluation (and, potentially, subsequent use of its findings), in which the operational test development context seemed likely to have a large impact on the feasibility of the project goals.

3.3 Phase 3 – Clarification of Evaluative Questions

In the second year of the project, a key goal of the evaluation was to clarify the evaluative questions. Towards this end, I asked the Advisory Panel to review evaluative questions as part of an early evaluation plan draft (fall 2008). Advisors provided written feedback that spanned: approaches to gathering summative evaluation evidence, key areas of focus, and ideas associated with the use of ECD-based structures – all helping to clarify evaluative questions for the project. In addition to input from the Advisors, a subset of project team members met (December 2008) to discuss our understandings of the values, interests, and expectations of key stakeholders (e.g., students, teachers, community members, science educators, assessment designers, NSF). This discussion of stakeholder concerns (see table in Appendix B) helped me to clarify and prioritize evaluative questions meeting critical information needs. The final (and working) set of evaluative questions (both summative and formative) is provided in Table 2.

Clarification of evaluative questions, including independent review by the Advisory Panel, and team-level consideration of stakeholder concerns, are important in contributing towards the NSF evaluation requirement that the evaluation plan, "poses significant questions that can be addressed empirically and that are central to the project's goals and objectives as well as contributing to understanding that meets current and expected educational demands of the nation on world-class criteria" (National Science Foundation, 2006). The careful vetting of the evaluation questions by the project team and nationally-recognized Advisory Panel members supported the emergence of the set of significant, mutually understood, and agreed-upon evaluation questions central to the project's goals and objectives. At this

phase of the project, less attention was given to how these questions might best be addressed, empirically.

Table 2. Evaluative Questions

Development of ECD-Based Structures
How are *design patterns* and other ECD-based structures best designed? How is this development carried out? To what extent do these structures explicate validity arguments for storyboards/ items? How are these structures used (if optional, are they used)? How well do the materials meet the needs of writers (users) for scaffolding writing/stimulating ideas? What do writers perceive as the "added value" of ECD-based materials? How efficacious are the ECD-based structures in storyboard and item writing?
Streamlining/Efficiencies
To what extent does applying the ECD conceptual framework/paradigm to the designing of a large-scale, state-level, assessment streamline and increase the efficiency of the test development process? What new structures and processes are created? In what ways are these incorporated into test development cycles? How is the PADI Design System extended in support of creating efficiencies? What technological benefits, if any, are associated with the creation and use of PADI structures for test design and presentation? Do these processes (applied to writer training and test development) save time and/or money? (Note that these questions will be addressed on the basis of information collected in the limited scale application of ECD-based ideas in MCA-II as are supported by the DR K-12 project)
Validity
To what extent does applying the ECD conceptual framework/paradigm to the designing of a large-scale, state-level, assessment increase the validity of test development processes? What new structures and processes are created? In what ways are these incorporated into test development cycles? How are key quality outcomes (e.g., validity of assessment arguments including theoretical rationale, item quality, test validity) impacted over time? Are there changes in item types or constructs? Is there impact on test accessibility and/or consequential validity? (Note that these questions will be addressed on the basis of information collected in the limited scale application of ECD-based ideas in MCA-II as are supported by the DR K-12 project)
Human Capacity
How is human capacity developed as a foundation to this work? How is storyboard/item writer training with ECD-based structures best designed? What other interstitial materials are needed for writer training to scaffolding decisions (e.g., about content, narrative, scenes, technical, art)? Does writers' engagement with ECD-based structures increase their perceived assessment literacy/capacity? In general, what types and kinds of expertise are created as a result of engagement in the project? For whom?
Dissemination
How applicable are project results to other statewide science assessment contexts? How are research and evaluative results disseminated to broader communities? What specific results are shared? What modes of dissemination are utilized (e.g., conferences, technical reports)? What audiences are reached? What is the impact of dissemination?

This process of articulating evaluative questions involves the areas of valuing and knowledge constuction. Clarification of the evaluative questions necessarily reflects what is valued in the project as well as what is unknown and worth pursuing. The first question that arises is, "who does the valuing?" In the case of this project, the valuing was done by the project team members (particularly, the PIs), the Advisory Panel members, and (indirectly) major stakeholder groups. The evaluator choose not to try to influence the values or articulation of questions, but rather, to serve to clarify those values and questions, and build consensus among the many contributing voices. Major values are reflected in the evaluation question;

these include focusing on validity evidence with respect to ECD-based structures, improving the test development process by increasing efficiency and validity, and increasing human capacity through engagement with the project and its ECD-based structures. Although different values might emerge on different development projects, these values reflect the promise of evidence-centered design as well as best practices in the field of assessment. The evaluative questions reflect the areas where stakeholders believe knowledge construction is most needed, tempered by the goals and feasibilities of this project. Questions with known answers are not asked (e.g., what should be the key attributes of *design patterns*), nor are questions for which empirical data is not pertinent or feasibly collected (e.g., a study comparing an ECD-based state-level test results with test results from another state).

3.4 Phase 4 – Design of the Evaluation Plan

3.4.1 General Evaluative Approach

The design of the evaluation plan was carried out in year 2 of the project and is based on the evaluative questions (synergistic with the project's goals). Towards this end, I asked the Advisory Panel to review and provide feedback on this evaluation plan (Fall 2008); their feedback was incorporated into the plan and is reflected in the remaining report. The evaluation plan was designed to meet evidence needs for formative and summative purposes, and along a number of major dimensions. It should be noted that formative evaluation[3] was given significant focus in this project, due to the iterative nature of the development and piloting of ECD-based resources. Table 3 provides general evaluative approaches applicable to a range of NSF-funded projects. For any given project, the evaluation plan should specify the amount of emphasis for each dimension and set of foci. In this project, the use of evaluative resources is a joint decision of the evaluator, project PIs, and Advisors.

Table 3. General Evaluative Approach Matrix for NSF Projects

Evaluative Dimension	**Formative Foci**	**Summative Foci**
Stakeholder Needs	**Solicit Stakeholders'** values, interests, expectations; clarify informational needs	**Share with Stakeholders** evidence-based results congruent with informational needs
Processes and Activities	**Communicate with Project Team** how the project might better implement processes and activities, and foster team members' understandings	**Report to NSF** the extent to which processes and activities are proceeding as planned

[3] In the RFP, NSF states, "Research and evaluation are likely to be formative in nature, providing information needed for the redesign of the resources, models, or technologies."

Goals and Outcomes	**Communicate with Project Team** the extent to which meeting goals, project strengths and weaknesses, project impacts	**Report to NSF** the extent to which the project is meeting its goals, strengths and weaknesses, project impacts

Stakeholder Needs

Stakeholders are any persons, groups, or populations potentially impacted by the project. For this project, these will include the project team and NSF, as well as other developers on the Minnesota Department of Education staff, storyboard and item writers, state-level administrators and legislators, educational practitioners, parents and children, and the educational research community. With respect to stakeholder needs, goals informing the evaluation plan included:

- thoroughly considering stakeholder needs pertaining to this project (see previous section, and Appendix B)
- using stakeholders' needs, interests, and expectations to inform this evaluation plan
- sharing evaluative results with stakeholders, congruent to their needs and interests

Processes and Activities

Processes and activities were specified in the project proposal, and are summarized in the logic model (see Figure 1). Goals informing the evaluation plan included:

- communicating with the project team, formatively and on an ongoing basis (through participation in weekly team meetings), how the project might better implement various processes and activities and foster team members' understandings
- measuring and reporting to NSF the extent to which project processes and activities are proceeding as planned

Goals and Outcomes

The project goals were articulated by the team and reviewed by the Advisors in Year 1 of the project (Table 1). Project outcomes, both proximal and distal, are specified in the logic model (Figure 2). Goals informing the evaluation plan included:

- determining, via qualitative and quantitative methods, the extent to which the project is meeting its goals – communicating these results to the project team (formatively) and to NSF (summatively)
- evaluating how the project might better meet its goals, sharing these judgments formatively with the project team
- determining the strengths, weaknesses, and impacts of the project, communicating these results to project stakeholders

3.4.2 Evaluative Methodology

The methodology for the evaluation will be discussed in terms of the areas of evaluative questions shown in Table 2. These are the development of ECD-based structures, streamlining and efficiencies, quality and validity, human capacity, and dissemination.

Development of ECD-Based Structures

ECD-based structures are those structures developed for the purposes of scaffolding writers' development of storyboards and items. These structures include *design patterns*, narrative structures (an attribute of *design patterns*), *templates*, and possibly Wizards. Development of these structures involves determining leverage points, alignment of the pertinent standards (Minnesota science standards, NSES) with the MCA-II benchmarks, developing the structures, and pilot-testing the structures.

The evaluation of the development of ECD-based structures involves observations of weekly team meetings, collection and analysis of iterative versions of the ECD-based structures, developers' articulations of the development processes and results, pilot testing of ECD-based structures, and writers' feedback on their experience with ECD-based structures. Weekly team meetings often involve the discussion, development, and critique of ECD-based structures; therefore, observation of these meetings provides evidence of the development process. The development of ECD-based structures, such as *design patterns*, is an iterative process; therefore, interim versions of particular structures are available and provide "snapshots" of the development process. Developers can provide their understandings of what is most salient in the processes of developing various structures. Structures can be pilot-tested with actual writers. Observations of the pilot testing were and can be made by the team members. Writers' feedback on the structures (e.g., via on-line surveys, focus groups, interviews) can include understandings, the usability, the "added value", and benefits of the structures (e.g., in terms of efficiency and validity).

Streamlining/Efficiencies

Streamlining and efficiencies have to do with ways that applying ECD to the designing of a large-scale, state-level assessment can streamline and increase the efficiency of that process. Since the application of ECD will be done in actual operational work for the MCA-II, but on a limited scale, the evaluation of efficiencies involves actual efficiencies created in the existing test development process, as well as potential efficiencies that would play out if ECD applications were scaled up — that is, if ECD were applied to the entire test development process (e.g., the development of all items and storyboards).

The evaluation of the streamlining and efficiencies afforded by ECD-based structures involves documenting the new structures and processes that are created, and how these are incorporated into the test development cycles through observation of project activities. In doing this, the evaluator will note if and how the PADI Design Systen is extended in creating these efficiencies (e.g., creation of new objects

such as new *design patterns*, creation of new classes such as a new structure for supporting efficiency in test development). Using operational item and storyboard timing data from 2007-2009 as a baseline, the evaluation plan calls for timing data in the context of operational use of ECD-based structures to be collected and analyzed to determine if the use of these structures reduce writing and/or reviewing time. Feedback from writers and users of these structures will be collected in terms of actual times to create and review products, efficiencies (e.g., perceived gains), and technological benefits (e.g., new, re-usable structures). In addition, feedback from the Advisory Panel will be solicited in terms of the potential scalability of newly created process supports and efficiencies.

Validity

In this project, we assume Messick's (1989) definition of validity as "the integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of infereces, based on scores." The project team will develop and provide design structures (e.g., *design patterns*) as resources to test developers, the intent of which is to open up more depth and breath of thinking in the design process. This incorporation of ECD is likely to have a lesser impact on the test development process itself (e.g., steps of the process) and a greater impact on writers' conceptualizations. Evidence of such changes in conceptualizations, or validity evidence, is likely to take two forms:

(1) articulation of assessment arguments
(2) validation indices (e.g., item quality, test validity)

The intention of the project is that writers' articulation of assessment arguments will be supported by design structures that provide some general aspects of assessment arguments. Writers, in interacting with these design structures, may change their conceptualizations of the assessment argument (e.g., think more deeply and broadly about theoretical rationale). Feedback from writers using these structures will be solicited in terms of their thinking in applying the design structures (e.g., via a think-aloud protocol) and perceived learning benefits (e.g., changes in articulation of assessment arguments; contribution of structure use to item and storyboard creation).

Finally, existing indices of item quality and test validity will be reviewed to determine if the application of ECD is, in any way, reducing the validity of resulting storyboards and items (test development outcomes). Where test developers used ECD-related tools to create particular products, characteristics of these products will be compared to similar products without the influence of ECD. Comparisons will be made on the basis of statistical indices (e.g., item slopes, reliabilities) and quality ratings (e.g., from content and bias panel reviews of storyboards and items). As part of this aspect of the evaluation, the project team will ask the Advisory Panel to review:

- the *design patterns* themselves, in terms of the extent to which they reflect the NSES and Minnesota standards and ground the generation of tasks
- how argument structures play out in grounding the assessment arguments behind particular tasks motivated by *desgin patterns*

Human Capacity

Human capacity includes understandings of the DR K-12 team and other test developers in terms of ECD-based ideas, the PADI Design System, and the MCA-II. The evaluation is concerned with how human capacity is developed, what types of expertise are shared, and what new expertise is created as a result of the project. Of particular interest is the design and results of writer training with respect to the ECD-based structures — what understandings about writer development in the context of ECD give rise to what training structures to what effect. Data sources to serve as evidence of human capacity are: (1) notes, observations, and transcriptions of team meetings, (2) documentation of writer training with ECD-based structures (pilot and operational settings), including training materials, (3) survey results (e.g., surveys of writers in terms of the usefulness and learning connected with particular ECD-derived tools), and (4) the DR K-12 technical report series.

From meeting notes and transcriptions, yearly summaries will be carried out in terms of types of human capacity shared, for whom, and in what ways. These will include any survey results. In terms of the technical report series, reports will be summarized in terms of the types and sources of expertise. Writers will be asked, for a given report, how that report extended their thinking. Writers and/or project PIs will be asked to document how each report is shared with others outside of our team (how much, how often, to what extent).

Dissemination

Dissemination will be measured in terms of points of contact between the DR K-12 project and broader audiences. Points of contact include:

- Presentations at professional conferences (e.g., CCSSO, AERA)
- Meetings with broader community members
- Technical reports
- Annual report

Conference presentations are documented in terms of: (1) presentation artifacts (e.g., PowerPoint presentations) and (2) on-line survey responses from presenters in terms of specific results shared, audiences that are reached, and perceived impact of the dissemination. In addition, for presentations that the evaluator attends, she will qualitatively assess the presentation impact through observation and conversation with audience members. Meetings with broader community members involving dissemination of results (e.g., state testing administrators) will be similarly documented in terms of

specific results shared, attendees, and perceived impact. Technical reports will be circulated via the internet; use will be documented as outlined in the previous section. The annual reports, written for NSF, will be retained as artifacts of the yearly activities and findings of the project, as well as a basis for potential results to be disseminated

Formative Feedback

Formative feedback of evaluative results will be given to the project team periodically, using both informal and formal mechanisms. Formally, evaluative results and recommendations will be shared with the project team on a yearly basis, concurrent with the NSF reporting schedule. Additionally, evaluative results will be shared at team meetings on an approximately quarterly basis. Informally, the external evaluator will attend and observe all team meetings (weekly) and provide input for shaping the project direction, based on evaluative findings.

NSF

The NSF requires that the external evaluator develop an evaluation plan that is designed to examine the extent to which the project has met its goals, reflective of the current literature and project context, addresses the development of materials and factors affecting implementation, that specifies evaluation questions, methods (appropriate to questions), data (existing, where possible), data analysis plans, and logical arguments for evidence-based conclusions, that specifies use of formative evaluation and how formative feedback will be communicated with the project team, and that contributes to understanding what factors contribute to project success. Further, they suggest that an Advisory Panel provide an independent review of this plan. The plan laid out in this section of the report is grounded in best evaluative practices, specifies critical evaluative questions and lays out methods for obtaining relevant evidence, specifies how formative feedback will be communicated with the project team, and is aligned with the logic model and operational constraints with an aim of contributing to understandings what factors lead to project success.

Dimensions of Evaluation

Developmemt of an evaluation plan is strongly associated with knowledge construction (Shadish, Cook, & Leviton, 1991), as well as valuing and evaluation practice. This evaluation plan is based on the epistemology of educational research, with a strong emphasis on observation, survey, and analysis of developmental artifacts. Although causal knowledge is not to be measured through a randomized, control-group type study, evidence from a variety of sources will be employed to determine the salient factors in the success of the study, as well as the likely causal impact of those factors. A critical aspect of the analysis of these data will be making claims and validly refuting rival hypotheses. At the phase of creating an evaluation plan, values are assumed from previous work defining evaluative areas, questions, and informational needs (e.g., Appendix A). Evaluative practice comes into play in determining the most

feasible and valid methods of data collection – often involving those members of the project team that are carrying out most of the work.

3.5 Phase 5 – Data Collection and Analysis

The data collection and analysis plan, and timeframes, are presented in this section. This work entails developing any needed evaluative instruments, such as surveys, engaging team members in instrument refinement and some aspects of data collection, collecting evaluative data, and analyzing the results. Formative results are communicated with the project team.

Evaluative Design and Planning

This work, previous described, involves developing the logic model and associated theory of change, drafting evaluative questions, responding to input from project team members and advisors, incorporating stakeholder concerns, and providing a draft and final evaluation plan to team members and Advisors. This work was completed within the first two years of the project (2007-2009)

Advisor Feedback and Communication

Communications and feedback from the Advisors include the previously described work with logic modeling, drafting the evaluation questions, and finalizing the evaluation plan. As an ongoing practice throughout the life of the project, I provide periodic evaluative newsletters to all Advisors, often soliciting their informal feedback on particular evaluative issues. Finally, a yearly Advisory Panel meeting is planned for the entire life of the project. Advisory feedback is given great consideration and focus in the evaluation work, impacting the evaluative decision-making in important ways.

Data Collection and Analysis: Development of ECD-Based Structures

Evaluative data collection on the development of ECD-based structures involves observing and documenting weekly team discussions of these developing structures, collecting iterative versions of the ECD-based structures; conducting interviews with developers of the structures in terms of the development process, challenges, and results; documenting the development of writer training for using these structures; pilot-testing the ECD-based structures; and obtaining writers' feedback on using the structures. Analyses of the iterative structures involves identifying changes to the structures, articulating considerations given to writers' needs for use of structures, summarizing developers' considerations in creating structures, and synthesizing input from writers themselves on use of structures.

Data Collection and Analysis: Streamlining and Efficiencies

Evaluative data collection for streamlining and efficiencies for the ECD-based structures involves observation and documentation of process supports and efficiencies, collection of timing data around use of process supports (e.g., item writing time with and without *design patterns*), writers' feedback: benefits,

perceived impacts, yearly analysis of efficiencies: what process supports, in what ways, to what benefit, and documentation of feedback from Advisory Panel on scalability of process supports and efficiencies. In terms of project goals, qualitative analyses of this corpus of data, as well as quantitative analyses of baseline versus treatment item and storyboard quality (e.g., writing time, quality criteria) are conducted as data become available.

Data Collection and Analysis: Validity

Evaluative data collection for validity of the ECD-based structures involves observation/ documentation of explication of assessment arguments (via representations, tools, think-alouds), collection of validity data for items and storyboards around use of argument structure supports (e.g., item slopes, reliabilities, quality ratings), feedback from writers on learning benefits and perceived impacts of structures, feedback from Advisory Panel on *design patterns*, and yearly analysis of validity: what structures, in what ways, to what benefit? In terms of project goals, qualitative analyses of this corpus of data, as well as quantitative analyses of baseline versus treatment item and storyboard validity, are conducted as data become available.

Data Collection and Analysis: Human Capacity

Evaluative data collection for human capacity on this project involves observations of weekly team meetings and other key meetings, materials from writer training sessions, writer surveys (e.g., Narrative Structures, *design patterns*), yearly analysis of human capacity development (e.g., from meetings, technical reports): what types, in what ways, for whom? Qualitative analyses of this corpus of data, in terms of project goals, are conducted as data become available.

Data Collection and Analysis: Dissemination

Evaluative data collection and analysis of dissemination efforts on this project involves collection of presentation artifacts (e.g., slides), on-line surveys of presenters, observations of some presentations, documentation of other meetings (e.g., NSF site visits), technical reports (available reports, available reports on-line, on-line hits, feedback from writers and PIs), and
yearly analysis of dissemination (presentations, technical reports, annual report). Qualitative analyses of this corpus of data, in terms of project goals, are conducted as data become available.

Evaluative Results

Evaluative results are shared via the following modes: evaluation meetings with DR K-12 Team (approximately quarterly), yearly evaluation report and recommendations (part of Annual Report), periodic presentations to DR K-12 team meetings, presentation of evaluative results at conferences/ meetings, and authoring or co-authoring technical reports. The next section will discuss communication and use of evaluative results.

NSF

The NSF requires that the evaluation plan address data collection and analysis plans, as well as logical arguments for evidence-based conclusions. The data collection and analysis plan laid out in this section of the report clearly specifies feasible and valid data sources, and capitalizes on existing data (e.g., baseline operational data, iterative versions of ECD-based structures). Data analysis plans for logical arguments for evidence-based conclusions are less explicitly laid out in this plan. Finally, in the case of a DR K-12 project, data sources can be triangulated for the needs of the evaluation as well as the research efforts.

Dimensions of Evaluation

This evaluation plan is based on the epistemology of educational research, with a strong emphasis on observation, survey, and analysis of developmental artifacts. Although the measurement of causal impact through a randomized controlled trial is not planned, evidence from a variety of sources will be employed to determine the salient factors in the success of the study, as well as the likely causal impact of those factors. A critical aspect of the analysis of these data will be making claims and validly refuting rival hypotheses. At the phase of planning data collection and analysis, values are assumed from previous work defining evaluative areas, questions, and informational needs (e.g., Appendix A). Evaluative practice comes into plan in working with team members and other stakeholders to provide timely informtion.

3.6 Phase 6 – Provision of Evaluative Information

In turn, evaluative results will be provided to key stakeholders in a timely manner, based on mutually agreed upon informational forms. The evaluator will communicate findings in a variety of ways. These include:

- Periodic presentations of evaluative findings or issues during weekly team meetings and conversations with team leaders.
- Communications with Advisors, such as periodic evaluation newsletters and informal phone calls
- Making yearly contribution of an evaluation report and recommendations to the NSF Annual Report
- Authoring and co-authoring technical reports with topics such as logic modeling and piloting results
- Authoring and co-authoring peer-reviewed journal articles
- Contributing to conference presentations (e.g., CCSSO, AERA), conference papers, and meetings

The NSF requires that, “All projects are expected to include an evaluation plan that examines the extent to which the project has met its goals…Summative components of evaluations must be conducted by a researcher or evaluator external to the project and submitted with the NSF final project report...For formative evaluation, plans should address how appropriate feedback will be given to the project leadership team so that it can make modifications to the project activities and address significant issues in the annual report.” Each year, the evaluation report (part of the NSF Annual Report) will provide summative information as to the extent to which the project is meeting each of its articulated goals. In addition, modes of communication of formative evaluative results (to the leadership team, and to other key stakeholders) are listed above; these include evaluative recommendations as part of the annual reporting process, periodic presentations to the project team, conversations with team leaders, and communications with Advisors. What constitutes appropriate feedback is a judgment on the part of the external evaluator. When I make such judgments and resulting communications, I rely on my understanding of the goals and values of the project, my knowledge of decision-makers goals and values (to motivate their engagement with formative recommendations), the evaluative evidence underlying such formative recommendations, and the potential impact of making a change on the basis of the formative recommendation.

The use of evaluative results is a cornerstone of any effective evaluation, especially for a project such as this one that is new in its development. Leviton and Hughes (1981) suggest that five conditions appear to effect the use of evaluative findings: (1) relevance, (2) communication between researchers and users, (3) information processing by users, (4) plausibility of research results, and (5) user involvement or advocacy. Fundamentally, the external evaluator must build and maintain professional relationships built on trust and mutual respect with all team members (especially decision-makers) to assure a strong foundation for communicating sensitive information. In advocating for use of evaluative findings, developing realistic, evidence-based recommendations helps assure relevance. Articulating such recommendations in a way that connects with the “prior knowledge” of the users (i.e., their frame work reference) helps support users information processing. Reminding decision-makers about the findings and recommendations keeps such information “on the front burner” in a way that can more effectively shape the direction of the project.

4.0 Summary

This technical report provided evaluation plans for the DR K-12 project, “Application of Evidence-Centered Design for State’s Large-Scale Science Assessment”. Evaluation work was presented and discussed in terms of evaluative requirements from the National Science Foundation, and in terms of five fundamental issues that undergird practical program evaluation: social programming, knowledge construction, valuing, knowledge use, and evaluation practice. The reports moved through six phases of evaluative work: (1) a logic modeling process, (2) definition of evaluative focus areas, (3) clarification of evaluative questions, (4) design of the evaluation plan, (5) collection and analysis of data, and (6) provision of evaluative information to stakeholders. Although evaluation plans lacked some forward thinking for particular analysis plans of collected data (evidence), rationales for evaluative design decision were provided and discussed at length, hopefully stimulating the thinking about process options for other, similar evaluation efforts.

References

Mark, M.M., Cooksy, L.J., & Trochim, W.M.K. (2009). Evaluation policy: An introduction and overview. *New Directions for Evaluation*, 123, 3-11.

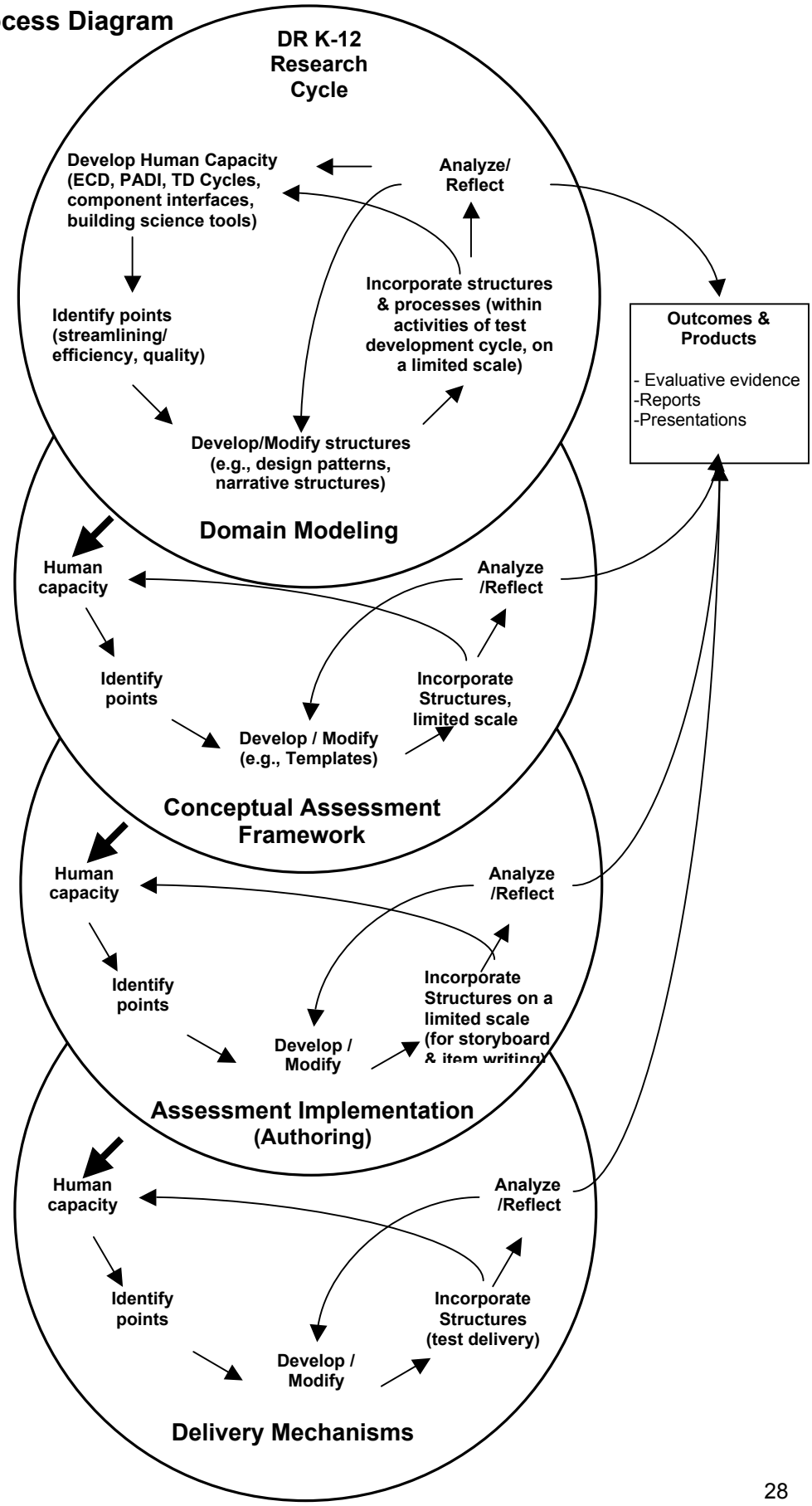
National Science Foundation (2006). DRK-12 Program Solicitation (06-593). http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf06593&org=NSF

Shadish, W.R., Cook, T.D., & Leviton, L.C. (1991). *Foundations of program evaluation.* Sage: Thousand Oaks, CA.

Skolits, G.J., Morrow, J.A, & Burr, E.M. (2009). Reconceptualizing evaluator roles. In *American Journal of Evaluation. Volume 30, No. 3.*

Weiss, C.H. (1998). *Evaluation*. (Second Edition). Prentice Hall: Upper Saddle River, NJ.

Appendix A: Process Diagram



Appendix B: Stakeholder Concerns

Key Audiences for Evaluation	Values, Interests, Expectations	Key Issues/ Evaluative Questions[4]
Our team (formative)	Pearson staff engagement with DPs (e.g., master designers) vs. writers [Dennis at nexus of writer training, DP use, items, storyboards] Learn about writers' process- what info held in mind as make decisions? Writer education – what are the goals and mechanisms?	How are DPs best designed? Interstitial materials needed - scaffolding decisions about content, narrative, scenes, technical, art, etc.? How to best design writer training? How efficacious are DPs? Does writers' engagement with DPs increase assessment literacy/capacity?
Pearson Staff, Pearson Writers	Improve products (validity), make designing/ writing more efficient Serving client (MDE) Improving writing capacity	Are DPs going to improve validity, efficiency? Are DPs worth the cost? Dev. of assessment literacy? Applicability to other SSAs? Sustainability/future support?
MDE Staff	A quality product for Minn. students – useful, accessible Reliable data, high validity Serves teachers, admins, legislators	**Validity: fidelity of measurement, technical qualities? Efficiency (saves money and time, or at least does not cost additional)
State-level Administrators and Legislators	Valid measure of science achievement Track trends across years Test product evidence for analysis of improvement (disaggregated) Comply with reporting requirements (NCLB, meeting peer review criteria)	Enhanced validity of the test to better serve schools, parents, students? Enhanced technical qualities of test? Does quality & fidelity of science achievement measurement improve? Bottom line - more efficient process?
District and School Administrators, Teachers	Teaching, learning Maximize student performance (AYP) Aligning, guiding instruction Accessibility of test to most students	Have item types, constructs changed? Impact on accessibility of test? Consequential validity – good info for decisions (AYP, instruction planning)?
Parents and Students	Parents – school choice (succeeding?) Students – not high stakes, but MDE individual score reports forthcoming	Parents – is test valid and useful? Students – validity, worth their time?
Edu. Research Community	*As articulated by our Advisory Group*	*See evaluation plan and Advisory feedback docs*
NSF (Julio Lopez-Ferrao)	*Formative evaluation, project goals, RFP*	*See project goals, eval questions*

[4] Cronbach suggests pursuing issues where the leverage is high (many stakeholders care), the uncertainty is high, and the yield is high (likely to find out something meaningful).

Sponsor
The National Science Foundation, Grant DRL - 0733172

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

Subgrantees
University of Maryland
Pearson