Managing complexity in classroom curriculum implementation sites: Triangulating Multi-Level Assessment and Evaluation

Jennifer K. Holbrook, Jackie Gray, Barbara B. Fasse, Paul J. Camp, Janet L. Kolodner

College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280
{holbrook, grayj, bfasse, pjcamp, jlk}@cc.gatech.edu

Abstract: This paper describes the efforts of the assessment and evaluation team of the Learning by Design™ Project. The project itself develops project-based middle school science units and ancillary materials and tools. The approach seeks to teach science content through design practice, but also to help teachers develop a classroom in which all members of the class are engaged in scientific practices and science culture. The assessment team's task is threefold: we seek to identify and develop tools and procedures to (1) understand the elements of classroom culture and teacher skills that are necessary for student success; (2) provide formative feedback for refining curriculum and support efforts for teachers; (3) deliver innovative assessment methods for use by teachers, students, and administrations that can be embedded within the everyday practices of the classroom and that can inform scaffolding efforts directly. Herein, we report on an assessment effort that spans multiple methods and seeks to answer multiple questions.

Context and Purpose of the LBD™ Assessment and Evaluation Effort:

Our project, Learning by Design™ (LBD™) attempts, like other cognitive-science-derived approaches to science education, to provide a learning environment that will foster the development of science knowledge and "habits of mind" for becoming self-regulated life-long learners. In our work on developing LBD™, we have focused on four major areas.

- **Curriculum development**: We are integrating the best of the relevant findings from the learning sciences to develop a curriculum approach that translates what we know about how people learn from the learning sciences and what we know about science education into a curriculum project.
- **Professional development**: We are making LBD™ work in real classrooms by providing the scaffolding for the teachers through both student and teacher materials. Our ongoing involvement with our collaborating teachers to refine our curriculum project allows us to

assess the actual needs of our learning communities and continue to develop scaffolding tools for a successful implementation.

- **Assessment:** We are measuring the effect of our implementations in a large metropolitan/suburban area by assessing student learning, focusing on assessing students' understanding of science and their ability to participate in the practices of science, design, planning, and teamwork. We are also developing embeddable rubrics and other tools for teachers and learners that allow them to assess their own progress in these areas.
- **Scaffolding development**: Scaffolding is help one gives a learner to allow them to perform slightly beyond her capability and such that she can learn from it. We view assessment and scaffolding as two sides of the same coin. The instantiation of this view harnesses the power one can get from a cycle of assessing and then providing the right kinds of scaffolding for learning. This requires that assessment of where students are be simple enough to be done often but powerful enough to allow student needs to be ascertained accurately. Assessment criteria and rubrics for assessment can provide an important component of scaffolding. This not only sets expectations for the students, but it also allows students to set reasonable short-term aims for themselves, especially if rubrics for assessment make clear the expected developmental progressions in acquisition of knowledge and/or complex skills.

What distinguishes assessment instruments from scaffolding instruments? Ideally, all assessment instruments serve to provide feedback for teachers and students so that appropriate scaffolding is provided to help a student or class bridge from level to level of concept understanding, skill acquisition, and metacognitive awareness. But to arrive at the point at which assessment and scaffolding are interwoven requires that each aspect of assessment be developed to the point that we understand what it can and can't tell us. Then the instrument must be integrated into the curriculum such that the assessment is easy to undertake, the results are easy to evaluate and distribute, and the outcome leads directly to the appropriate level of scaffolding. Meanwhile, data from assessment instruments serve program evaluation needs as well. In designing instruments that serve for formative and program evaluation for the LBD™ project, we generally start by developing instruments and methodology to be used external to the curriculum for program evaluation purposes. But each was developed with the purpose of eventual integration in the curriculum as tools for use by teachers and students, rather than by

external assessors. In this paper, we explain how each instrument and related rubrics have been developed, how it has been used in program evaluation, what we have been finding through these instruments, and how we envision it being integrated into the curriculum.

Making assessment information available to students can encourage self-directed learning in students, as it provides them with feedback on their progress. Assessment is also essential to providing instructional intervention or support for the learner (scaffolding). Supporting and extending the developing understandings that students exhibit is central to our approach. Our assessment tools reveal this developing understanding which, in turn, is used to determine the application of the most relevant scaffolding. Students vary in their understanding and developmental progressions towards knowledge and skill acquisition. The assessment and scaffolding tools we have developed provide an important practical translation of what we know from how to assess and scaffold complex cognitive activity.

Our approach is distinguished by (i) our focus on assessing and scaffolding science and collaborative practices (where others focus on assessing knowledge), (ii) our development of resources that scaffold students' performance and learning in a project-based classroom and at the same time scaffold teachers as they develop facilitation skills, and (iii) our design of assessment and scaffolding tools in collaboration with teachers and students so that we know we are addressing real issues in actual classrooms. Our aim in addressing these goals is for students to be able to learn content in such a way that they can apply it in a variety of situations (promote transfer) and to become skilled science learners, becoming competent at participating in the practices of science and science learning.

Our assessment goals have been to provide alternative, more dynamic ways to assess student learning and skill mastery beyond static measures like standardized achievement tests; involve our teachers as collaborators in each step of this developmental process; and validate those processes and tools that would have value to other teachers of middle school students. A lesson we continue to acknowledge as we work with our teachers is that we all want students who have high achievement tests scores, but we also want students who are learning the deep principles and processes of science that will go on to become the innovators, designers and problem solvers of tomorrow. We want to extend what counts as learning in science to include the process skills that will equip our students to learn life long.

Indeed, the science education community and the recently-published American standards

about science literacy want students to gain both competencies just listed (American Association for the Advancement of Science (AAAS), 1993) – to learn science concepts in a way that allows them to apply those concepts to new situations as they arise and to become enculturated into the practices of scientists. This includes being able to enact science process skills such as inquiry, observation, measurement, experiment design, modeling, informed decision making, and communication of results, as well as social practices such as learning from each other. The aim is for students to learn in ways that will allow them to take part in those practices in skilled ways both inside and outside the classroom Transfer, in the context of science learning, means gaining expertise in the practices of the scientific community (e.g., designing experiments, managing variables, justifying with evidence, analyzing results, planning investigations, communicating ideas, communicating results, incrementally building up one's understanding of a new concept) and learning scientific concepts and their conditions of applicability in order to engage in scientific reasoning.

From several iterations on the development of our assessment and scaffolding tools, and from the data we've collected to show student learning with these tools, we have had an impact on addressing the following national standards for middle school students in science:.

- Design an investigation
- Communicate investigative process
- Answer original question
- Communicate results
- Understand relationship between explanation and evidence
- Use tools to gather, interpret and analyze data (including mathematics and computers)
- Understand explanation, in contrast with mere description
- Consider alternative explanations presented to them (not necessarily ones they themselves come up with)

The assessment and scaffolding tools we've designed address these standards and more. Our results so far show that the scaffolding tools and assessment instruments we have designed have potential for impacting the development of these very important skills, routinely practiced in our LBD™ classrooms. We also note from the standards (AAAS, 1993) several practices that students tend to have difficulty with, and LBD provides extensive practice opportunities and scaffolding for those:

- Identifying variables
- Controlling for more than one variable
- Understanding that different variables have different levels of effect (lots, little, none)
- Mentioning evidence against their conclusions
- Having conclusions that differ from previous opinion
- Coming up with alternative explanations.

LBD units address specific science content as well, and our units actually cover more content and process skill development than these standards. For example, our students develop great skill in collaboration. The list does, at least, highlight the kinds of things we're reciprocally assessing and scaffolding. Our assessment and scaffolding tools are developed to assess and support the acquisition of these skills and "habits of mind".

Program-Evaluation-Specific Methods and Instruments

1. Ethnography

Evaluation of the earliest LBD™ curriculum implementations focused on providing formative feedback for the curriculum development effort. For the most part, this called for ethnographic methodology—case studies of a small number of students and the environment in which they were operating (the unit, the students' expectations of the classroom culture, the teacher's approach, the social environment of the class and school, the school's values and expectations, etc.) These case studies included frequent passive and participant observation and taped and informal interviews. Such feedback helped identify the differences between what the unit developers assumed about the classroom culture in which the units would be instantiated, the actual environment in which the units were implemented.

After several major revisions, the developers were fairly satisfied with the unit format and activities, and with the type of training offered to teachers. As this happened, the purpose of classroom observation evolved: now we were interested in studying how the variations in implementation affected the success of the curriculum. "Success" in this case is broadly defined as how well the students learn target content, science and design practices, collaborative skills, and metacognitive strategies; also, "success" includes the students' level of engagement with and enthusiasm for the curriculum, and their teacher's satisfaction with teaching the curriculum.

Operationalizing and measuring such definitions of success led to the use of very different methods of assessment, detailed below. However, it was still through ethnographic methods that local variation continues to be analyzed. Now, the ethnographic efforts focus on understanding how different teachers with different teaching styles, knowledge of the target science content, and knowledge of science and design practice, make the affordances of LBD available to students. Also, the ethnography gives us information about how students are responding, their levels of engagement, what's difficult for students, how students use one another as resources, and so on. From this, we are learning about the affordances our materials and ancillary training provide for the students and the teachers, and what is still needed.

In the first years of LBD, when only 1-4 teachers were implementing our units at a time, we were able to spend enough time with each teacher to do in-depth case studies of all. In the last few years, however, the number of implementing teachers per year (and the geographic area over which they are spread) has grown as our staffing has stayed constant. It was therefore necessary to develop methods of assessment that could be reliably used by amateurs, which would yield valid data about the differences in:

- the fidelity of unit implementation;
- the skill with which teachers orchestrated an inquiry-based learning,
- the engagement of students with the project.

Thus, we create our ethnography from data gathered through four strategies (Fasse & Kolodner, 2000). First, we've developed two observation instruments to help observers focus their observations in all of the classrooms. While this flies in the face of qualitative methodology, we do have a practical need to make sure that our untrained observers include the taken-for-granted world in their notes. We target a subset of teachers to visit at least once a week. This subset, is chosen based on specific questions we need to answer. For example, teacher was included in the set because of both strong science background and specific training in teaching inquiry-based classes, while another was included because she was teaching science for the first time ever. This allows us to understand what the teacher needs to know to make the implementation successful; in turn, this information is used in developing teacher support and in helping teachers decide if the curriculum is appropriate for them to use. Second, we interleave thick description (Geertz, 1983) from our observations with description derived from video documentary. We videotape all of the LBD™ and the comparison classrooms at least twice

during the year; some are selected for much more frequent videotaping. Third, we meet with our teachers in focus groups every six weeks to learn what works and doesn't work in their classrooms and to allow them to share their experiences with each other. Fourth, we have team members use prompt sheets and checklists to create a description of classes each time they have reason to visit.

The data's audit trail includes field notes from observations and interviews, written evaluations of class visits based on the checklist and prompt instruments, transcriptions of audiotapes, and written summaries and coding of videotapes (Lincoln & Guba, 1985; Spradley, 1980). Our staff ethnographer organizes and maintains these data, and then members of the team make use of whichever resource is most appropriate to inform specific research questions as needed.

2.--Measuring LBD™ Fidelity of Implementation and Inquiry-Friendly Practices in LBD™ and Comparison Classrooms

The teachers who implement LBD™ represent a wide variety of teaching styles, knowledge of target science, experience in science methodology, and experience with teaching and exposure to inquiry-based teaching methods. Then too, the classes and schools in which they teach vary a great deal—there are economic disparities between schools and districts and disparities of administration style; there are class cohorts that have a wide variety of student abilities vs. class cohorts that have been restricted, sometimes to honors students, perhaps to at-risk students. Each of these features will affect how the teacher implements the unit, from the amount of time taken on a given construction activity to the types of questions the teacher frames for class discussion, from decisions made about grading collaborative efforts to the emphasis placed on identifying methodological variation in experiments.

Learning by Design™ units provide teachers and students with specific sequences of activities, group and individual assignments, and carefully articulated expectations. The ways in which teachers depart from the unit and the reasons for these departures must be documented so that, when faced with different outcomes among different teachers' classes, we are able to identify covariates of the unit.

As discussed above, the thick description, interviews, and videotapes that we gather from some classes make it easy to gauge the fidelity of unit implementation. However, in classes that

are not targeted for intense ethnography, it is still important to gather an accurate description of the fidelity of implementation.

Staff visits to classrooms afford the opportunity to get a "snapshot" of implementation fidelity. However, unless the staff member who is visiting is well versed in qualitative methodology, it is difficult to know what counts as evidence in forming opinions and conclusions of implementation fidelity. It is also important that any such summaries of implementation fidelity meet a standard of objectivity, or at least interobserver reliability.

The Observation Prompt Tool (Holbrook, Gray & Fasse, 1999).and LBD™ Fidelity Report Card are both designed specifically for use with our evaluation effort. The Observation Prompt Tool prompts for information about which types of LBD™ activities take place during the visit, and for specific aspects of how such activities are carried out, as well as a description of the environment in which the visit is carried out. It also prompts for descriptions of teacher and student roles during the various activities, and for descriptions of how well these roles are carried out. For example, in a discussion, the teacher may adopt the role of lecturer or inquisitor, or may moderate a discussion among the students. The students may actively seek to question one another, or they may expect the teacher to provide the discussion framework. The questions themselves may be more- or less-well-suited to promoting connections between class actions and science concepts Each of these areas has a set of set of several statements and questions to cue the observer to describe specific elements of an activity, and to speak to specific issues about the classroom.

The form of data gathered from such visits includes a written summary of the visit and a completed copy of the OPT form; it often also includes a videotape of the class. The written summary is indexed to the appropriate sections of the OPT, so that the evidence and context for a specific descriptive choice is preserved.

The Fidelity Report Card is intended to be used to sum up a set of such observations. Evaluative statements for both students and teachers are included. The items are scored on a 5 point scale (Unsatisfactory—Needs much improvement—Meets fair expectations—Good—ideal). Many of the items could be used to evaluate any science classroom (e.g., "Students use science vocabulary and/ measurement techniques w/accuracy"; "Students listen/discuss/consider ideas/suggestions of others w/in group"; "Teacher knowledge of the specific science area"; "Teacher shows flexibility for changing plans when indicated by student needs". Other

questions are specific to LBD™, such as: 'Students independently employ LBD rituals to inform decisions"; Teacher uses the [LBD] rituals as scaffolding tools to promote or model the process of reflection". When data across time are available, the Fidelity Report Card is applied separately to early and subsequent time frames to see how the class culture evolves throughout the year.

What results is a Fidelity of LBD™ Implementation score (or set of scores, when employed across time) for each teacher in the program. These scores are ranked ordinally, and the scale is used as a covariate in comparing results on written tests, performance assessments, and structured interviews. Three staff members each score the teachers for Fidelity of Implementation, based on the OPT and written reports.

We obtain high reliability ($r > .9$) on all items for Fidelity of Implementation scores based on reports from at least halfway through the content in the school year. The reliability for Fidelity of Implementation scores based on reports from earlier in the school year are less reliable the earlier the time frame. We are currently investigating whether this is attributable to some scorers setting developmentally-relative standards and other scorers, absolute standards.

LBD™ specifically seeks to promote an inquiry-friendly culture through the activities, the written materials, the assignments, and the in-service teacher training and support. We believe that such a culture is key to promoting deep learning. Thus, it is important to document not only the differences in the written materials and activities of LBD™ and comparison classes, but the differences in classroom culture, as part of the context in which outcomes for LBD and comparison classes are discussed.

We approach documenting inquiry-friendly culture in the same general way that we approach documenting fidelity of implementation, i.e., employing those sections of the OPT which are not curriculum-specific, and a subscale of the Fidelity Report Card that focuses on inquiry skills only. The scores on this subscale for both LBD™ and comparison teachers are then ranked ordinally and used as a covariate in comparing results on written tests, performance assessments, and structured interviews.

The value of ranking teachers' classes for "inquiry friendliness" and "fidelity of implementation" becomes clear as we begin to look at measures of learning outcomes. Here, from Holbrook, Fasse, and Camp (in preparation) are some examples:

We find that the teacher who "used science terminology well" (Fidelity Report Card item 33) had students who did the same (FRC item 1), and those same students have significantly higher scores in performance assessments and structured interviews. The teacher who used design terminology and orchestrated LBD™ rituals well, but did not have great science knowledge, (FRC items 32a, 26) had students who spontaneously used the rituals to solve their own problems (FRC item 15) and transferred the language and methodology both in performance assessments and to science fair projects (Fidelity Report Card items 2,3,4, 16), but had only mildly significant pre- to post-test gains on the target science content. The comparison teacher who used science terminology very well (FRC item 33) but did not identify science concepts as they emerged in student discussion, questions, and demonstrations (FRC item 41), and did not identify science in everyday events or cases (FRC item 42) had students who did not use science terminology well (FRC item 1) and who did not show a significant change from pre-to post-test on content knowledge.

And so on. In other words, we are beginning to be able to tie specific teacher and student habits in the classroom to outcomes on the measures of target science content learning and science practice, in both LBD and non-LBD classrooms. These predictors in turn help us to focus our ethnography more tightly, and helps us to tailor teacher training and curricular materials.

3.—Targeting Specific Teacher Practices

As the curriculum units changed less from implementation to implementation, we were able to make preliminary guesses, based on ethnography and outcome data, about how particular practices were most closely related to various aspects of student success. By narrowing our scope to studying two practices in depth per implementation year, we added a new dimension to the ethnographic data being gathered. This year, for example, we have looked at how teachers make use of the practice of articulating Design Rules of Thumb (Crismond, Camp, & Ryan, 2001). The idea behind these rules is that, as class groups design an artifact and conduct tests upon it, their data will help them form conclusions about the intrinsic constraints of the task. Articulating these constraints can lead from the specific artifact's behavior in a specific set of circumstances to the recognition of an underlying aspect of science. For example, student groups seek to develop the best balloon-propelled car by testing a number of variables, including the aperture of the straw through which air escapes from the balloon. The "best" width is the width

that allows the most air to be released—up to a point. The student group may begin by articulating the design rule as a piece of advice about straw width. Sometimes, the advice is too artifact-specific (e.g., "use straws of 7/8' diameter). Teachers who push for reasons to back up the advice are requiring the students to back up their conclusions with evidence. Teachers who understand the science will orchestrate the discussion moving, first toward generalizing the rule, (e.g., "use wide straws instead of narrow ones") then push toward generalizing of why the wider straw is better, and on to consider about whether there's a ratio of balloon size to aperture. They will also help the students remember that air is a gas, and that it's being released through a rigid tube; they may have students consider the behavior of the balloon and the behavior of a rocket. Thus, when the rule of thumb is articulated, the science underlying the concept is much better understood, and the opportunity for far transfer in applying the rule of thumb is enhanced.

4. Measuring student outcomes

Among the questions we seek to answer through our program evaluation are:

(1) Do students in LBD classes learn more of the target content than students in comparable

classes that use more traditional curricula, teaching methods, and culture?

(2) How great is the expertise developed in the target content areas by LBD students and their comparison counterparts?

(3) Do LBD students gain greater ancillary skills that suggest they have "learned how to learn" better than their comparison counterparts? In particular:

- do LBD students give more detailed explanations of their answers (in prompted situations, or spontaneously)?
- do LBD students show greater ability to transfer knowledge?
- do LBD students reference cases more often?
- do LBD students understand scientific methodology issues better?
- do LBD students show greater facility in reading and/or generating graphs, tables, illustrations?[1]

Quantitative assessment techniques must be used in ascertaining these outcomes, for credibility in the larger community comes from being able to compare outcomes on agreed-upon

measures. However, much of what passes for knowledge assessment is justifiably criticized as favoring rote memorization without understanding (e.g., Wiggins, 1993). The difficulty for those who design assessment instruments is to find a way that measures knowledge objectively, reliably, and validly without measuring knowledge at too shallow a level. Fortunately, alternative assessment methods are becoming more widely available and accepted, and we have been able to adapt numerous available instruments and techniques to meet our own program evaluation needs.

Recognition—Based Content Tests

In developing the current versions of content tests for earth science and physical science, we sought to keep the tests simple to administer, but at the same time, to include questions that allowed a richer understanding of the depth of understanding of target science content. The content tests are in the familiar multiple-choice question format, but the question scenarios and the choices themselves are carefully crafted to allow analysis of the stage of understanding that the student is at before instruction, and what qualitative differences in understanding occur as a result of instruction.

Question Origins

Two types of questions were specifically included to help answer the questions about student outcomes posed above.

One type, which we call "Explanatory Answer," links a multiple-choice response with a required explanation of the response choice. The score on the question is based on giving the correct response among the choices and explaining why this is the correct response.

Another type of question is multiple-choice in nature, but the choices are carefully crafted to reveal the depth of understanding of the concept at issue, a methodology Thornton (1997) refers to as "conceptual dynamics.". These questions were published items in the Tools for Scientific Thinking Force and Motion Conceptual Evaluation, or FMCE,(Sokoloff & Thornton, 1989) and the Force Concept Inventory, or FCI (Hestenes, Wells, & Swackhamer 1992) which were developed to track the development of depth of understanding of a number of physical science concepts. These tests were designed to be administered to secondary and college

students. The questions that we selected were adapted with simpler wording and choice selections for the younger students who are being assessed in the LBD project.

Thirteen of the questions on the exam were adapted from for study of the development of depth of understanding of concepts of force and motion. Many of these questions were clustered together to probe understanding of a specific concept carefully. Questions 14-16 look at the effect of force on a given velocity; questions 17-18 probe the understanding of acceleration as a change in direction; questions 20-24 are about force, mass, acceleration, and "equal & opposite reactions", and questions 25-27 are specifically about gravity as a force in acceleration on a hill.

In comparing LBD classes to comparison classes, we find significant Curriculum X Test Administration Time interaction for learning gains on target science content tests (Holbrook, Gray, Camp, & Fasse, in preparation.) The gains for teachers with little previous experience in facilitating project-based, inquiry-friendly classrooms and with modest physical science knowledge are only modest (a mean gain of about 1/2 pt. on the target science items), though significantly greater ($p < .05$) than those of the matched comparison teacher (no mean gain on the target science items,

Another pairing is more interesting: two teachers who both had honors classes, and were both certified for physical science teaching, and were both experienced in facilitating project-based, inquiry-friendly classrooms. The LBD™ teacher's mean pre-test scores on target science content were similar to those of the matched comparison teacher, but whereas the LBD teacher's classes more than doubled (from a mean of about 3.2 to a mean of 8), the comparison teacher's class only gained about 20% (from a mean of about 3.2 to a mean of about 4.5). The Curriculum X Test Administration Time interaction for learning gains for this LBD™ set of classes was significant $p <.001$). These results, taken together, seem to show how important teacher knowledge and experience with inquiry-rich teaching methods are to help students make meaningful knowledge gains. Yet the curriculum also has an important role to play.

The number of comparisons is still too small to make strong claims in this area; as we go to press, a larger set of systematic comparisons is underway.

Coding "Depth Of Understanding" Questions

The content test is administered both pre- and post-curriculum. One way that we compare LBD™ and comparison student learning of target science is to compare LBD and

comparison student's overall change of scores on the test items focussing on target content. For such an analysis, it is easiest to interpret findings if questions are coded as being "correct" or "incorrect" as on traditional tests, rather than trying to encode different levels of understanding as interval values. Differing stages of understanding could not appropriately be interpreted on an interval scale, as is assumed by the types of repeated-measures analyses we intend to employ. However, providing nominal and ordinal codes, then using descriptive and non-parametric analyses on these items specifically, will allow us to study a number of changes.

The coding scheme we devised, then, serves a dual purpose: (1) to provide simple "correct/incorrect" labels of most items[2] on the test, (2) to provide more detailed labels that could serve to compare how pre-curriculum and post-curriculum misconceptions and missing knowledge differ from one another. Codes were assigned as follows: On multiple choice questions, the letter of the multiple choice answer was preserved in one stage of coding. In another stage, the letters were matched against and answer key and converted into "correct" and "incorrect" answers: 1 was given for a correct answer; 0 for an incorrect answer.

On explanatory answers questions, incorrect answers were given additional coding labels of a single letter hat described the rationale of the answer. For example, "m" on question 27 might stand for "mass". Coding labels for each explanatory answer were unique to that question (thus, "m" might stand for "measure" in question 7). As these labels do not affect the designation of correct v. incorrect in the repeated measures analysis, and each question's answer pattern must be studied separately, there is no need for these codes to be systematic from question to question, and might in fact make it necessary to provide longer codes or use numeric codes which are less mnemonic for coder and analyst alike.

Analyzing for Nature of Concept Change

For the questions from the FCI and FCME we've looked at the differences in answer patterns from pre- to post tests. The concept inventory questions were originally designed to tease apart different levels of concept development. Thus, even answers that are designated "wrong" can show movement from one level of understanding (e.g.., guessing by using contextual clues, Aristotelean-level concepts) to another (partial understanding, with remaining

[2] A few items are scored as full, partial, or no-credit items.

misconceptions). Should the patterns of responses differ between LBD™ and comparison students, we will be able to compare their relative depth of understanding of the concepts.

The first step was to model question-answer patterns that would be indicative of each level of understanding. Some of this modeling work was actually done in the devising of the questions. However, guessing patterns were not modeled by the test authors. Random guessing would of course be indicated by relatively even distribution of answers. We also model a more sophisticated guessing strategy that will take each question's contextual statements as the basis for the answer, rather than the physical law which is to be inferred by the question explication.

Each model predicted a specific pattern of response preferences for each question. We ranked these models from least-to-greatest concept understanding, with more specific guessing strategies ranked higher than less sophisticated guessing strategies.

Current Work:

Similarly, on the explanatory answer questions, we are looking at the differences in both the *quantity* of correct choice/explanation pairs from pre-test to post-test, and the nature of the explanations, especially incorrect explanations. We expect to find (1) on the pre-test, many of the explanations will indicate guessing (answers such as "I just picked one" are coded as guesses), and that there will be a large number of students who do not attempt any explanation at all. (2) Incorrect choices and explanations will indicate a later developmental stage of the concept on the post-test than the pre-test.

Additionally, with explanations, we can look at the *nature* of explanation. We can see whether the answer cites aspects of the question, physical science/earth science principles, or cases from personal/class experience. Differences between LBD and comparison students on the nature of explanation from pre-test to post-test will allow us insight into most of the questions cited under 3 at the beginning of this document. (The exceptions are noted in the footnote.)

Finally, we will be able to see whether LBD students are more likely than comparison students simply to articulate explanations on their questions from pre-test to post-test, for among the encoded responses is "no answer." In a very perfunctory scan of the complete answer worksheet we see a large difference between the number of pre-test and post-test "no answer" codes. To show that LBD students have developed a habit of explaining their answers, even when those answers are tenuously grasped, would support our claim that LBD classrooms

provide and environment in which both "thinking through" and explanation are valued by students. The first step in resolving misconceptions is being able to identify them; students' ability to articulate a conception helps both them and the teacher in such identification.

Performance Assessments and rubrics

Goals: (1) To develop an instrument/methodology that measures both content knowledge and science process skills, such that the dynamic aspects of the thinking processes necessary to transfer the knowledge and apply the skills is measurable.

(2) To develop an instrument/methodology that allows assessment of students' ability to use evidence to answer questions and in support of assertions or recommendations.

(3) To integrate the instrument/methodology into the curriculum such that (a) assessment is easy to accomplish as a natural part of the curriculum, and (b) scoring rubrics are learnable, reliable, easy to use, and easy to explain to students.

It is difficult to actually measure complex cognitive activity and content tests, since traditional measures do not generally offer a window into how complex skills actually play out or are developed. Zimmerman (2000) reviews the literature on scientific reasoning and makes the point that to assess scientific reasoning, activities must be examined during which such reasoning is needed. Performance assessment tasks are just that. They allow reasoning skills to be examined in knowledge-rich contexts.

We have gained expertise in the past two years in designing and adapting performance tasks that can be used to assess student learning of skills and practices and in creating rubrics that can be used to analyze the extent to which students are participating, and we've developed several such tasks and the materials needed to use them reliably for assessment. Preliminary evidence based on these tasks shows that LBD™'s practices promote transfer in the subset of the students we have evaluated, showing us that such tasks can be used for assessment of skills learning and can be coded reliably.

We have adapted performance assessment tasks to follow a format that allows us to better assess the collaboration and science process skills that we seek to promote in the LBD™ curricula. The task is designed in three parts: (I) students *design an experiment* to gather

evidence to address an issue in the context of a real-world problem; (ii) students work in groups to run a specified experiment with materials we have provided, and gather data from this experiment; (iii) students answer questions that require them to utilize the data they gathered, and to apply their knowledge of science to interpret the data. The quasi-experimental design has different classes assigned to different participation conditions: Some classes have the students do all three parts of the task as a group, writing a single group answer; some classes have the students run the experiment as a group, but to work as individuals on parts 1 (designing/writing an experiment) and 3 (interpreting data, answering questions); and some classes have the students work together on all three parts to develop answers, but each student writes these answers in his/her own words.

We videotape the two conditions in which groups of students work together throughout the task. The design-an-experiment part of the task allows us opportunity to judge group ability to design an investigation, their understanding of what a variable is, and their ability to control variables, among other things. The middle part helps us determine their ability to carry out a procedure carefully and correctly: to measure, observe, and record. The third part allows us to determine if they know how to use evidence to justify and how well they can explain. All three parts provide evidence about their collaboration and communication capabilities and their facility at remembering and applying important classroom lessons.

An example task may help bring this to life. In "Where the Rubber Meets the Road," adapted from a task developed by the Kentucky Department of Education and now available through the PALS Performance Assessment Links in Science website (PALS 1999). Part I has students design an experiment that compares the efficacy of two tire types that differ in the hardness of the rubber used when tested in different road conditions. The science concept being tested is understanding of the force needed to overcome sliding friction.

Coding categories include negotiations during collaboration; distribution of the task; use of prior knowledge; adequacy of prior knowledge mentioned; science talk; science practice; and self checks during the design of the experiment, and each group is scored on a likert scale of 1 - 5, with 5 being the highest score. (See Appendix 1 for examples from the coding scheme developed to assess collaboration and science practice skills during these tasks.).

Preliminary analyses show that in general, LBD™ students show greater collaborative skills and use more sophisticated science practice than their comparison counterparts (Gray, Camp, Holbrook, & Fasse, 2001).

Structured Interviews

While performance assessments allow for assessment of dynamic processes, they have some disadvantages for use in the curriculum:

(i) they take a large chunk of time to implement, evaluate, and provide feedback

(ii) the written-answer format does not allow for on-line probing of understanding or skills (iii) they take the form of a class activity that is separate from the class project, so they are most naturally treated as tests

(iv) they are intended to be implemented for the whole class

In response to these concerns, we have been developing an assessment method to meet the same goals, but that allows for more flexibility in implementation. Structured interviews can be used for individual or small group assessment by a teacher (or assessment team member). They can be administered in multiple ways: by having a student read a scenario or run a simulation or devise a solution to an orally-presented problem. They can be administered one-on-one and casually, pulling a student who is at leisure aside for a few minutes, or they can be assigned to a group or a class as a quiz. The salient feature is that the student is to describe the way s/he will go about solving a problem. The student(s) then carries out their proposed methodology, and reports on results. The salient feature of the structured interview is the interactive nature; the assessor is not simply recording the student's responses, but using probing questions, observing the procedure, asking questions about choices and results, and noting what the student spontaneously offers and what the student is able to do with various levels of scaffolding.

This year marks the pilot effort of structured interviews. We have used several different tasks with subsets of students, and we have begun refining the interview framework for each. We intend these to eventually be integrated into the curriculum, having teachers use them for either informal or formal assessment.

Student self assessment survey

We have collected an exhaustive list of skills for teachers and students to use to guide student progress throughout our LBD units, along with questions that students can ask themselves about each in order to assess their capabilities and developmental progress. We then targeted two specific aspects of learning for self assessment, collaboration skills and active research skills. Twenty survey items were selected and adapted to cover these two topics. The Twenty Questions survey is intended to be used after each major activity. It is administered after each of the launcher unit activities, then every few weeks as specific sub-challenges to the overall problem are completed. The students keep the survey with their notebook; they have access to their own survey answers each time they look at the survey, and the instructions encourage them to notice differences from administration to administration.

The Twenty Questions survey was administered multiple times in a subset of our classrooms last year. We were able to analyze the data from four of our teachers' classes. We had different predictions for the two aspects that were being self-assessed. For the collaborative skills, we predicted that (1) students would rate themselves highly in the beginning of the year, (2) would show lower ratings of themselves around midpoint since they would have had actual experience working as teams, collaborating, doing research and engaging in challenging design problems, and (3) would rate themselves more highly again at the end of the units, reflecting a more accurate picture of their skill development. This U-shaped developmental progression was found to be significant on several items from the survey. Students in one teachers' four classes (N= 120 students) showed significant change in their ability to think about the function of the objects/artifacts they design; identify and examine alternative solutions to problems; and, make decisions based on evidence and not just guesses or opinions. (see Gray & Holbrook, in preparation, for the data report on all teachers participating in this evaluation effort). We are encouraged by these predicted significant outcomes.

For the active research skills self-assessment segment, we predicted that students would tend to score lower at the beginning of the year and that the scores would get higher over time. Our reasoning is that the students would tend to judge their ability to know where to turn for information and how and when to design experiments would be low; as they practices these skills, their confidence should grow. We are currently analyzing the results.

Future Work: Embedding Assessment for Program Evaluation and Class Scaffolding

In the introduction, we said that one of the goals of the LBD™ project is to develop science units based on what we know about how students will learn and retain that learning. One of the most important aspects of helping students learn is to provide appropriate scaffolding so that students are able to form bridges from less-sophisticated to more sophisticated understanding. But to form such bridges, we must have ways of ascertaining where the student is beginning, as well as knowing where we wish the student to end up. And traditionally, classroom assessment is a capstone experience—ascertaining where the student ended up.

Among the criteria we have used in designing our program evaluation and research tools, we have always included the need to find ways to adapt assessment techniques to be used by teachers, administrators and students themselves as part of the classroom experience. By finding ways to embed assessment in the curriculum itself, we seek to provide teachers with an understanding of what students still struggle with, and also what students may be able to bring to bear from related experiences.

The first step, of course, has been to develop and test instruments and rubrics, and this dovetailed with the requirements of the program evaluation effort. Then, we consider what we intend for the curriculum to scaffold, and what instrument might be useful to the class to assess the specific needs. Adaptation and adoption of the various instruments described herein is at a very early stage. We have included the Twenty Questions Survey as part of the teacher materials, and intend to embed the survey into student materials in a future unit revision. Currently, the teachers administer the survey only for program evaluation purposes, but we have begun developing ways for teachers to employ the surveys directly in their assessment of class needs.

We have also been working on methods to integrate performance assessment tasks into the class units. Problem-based curricula are natural settings for assessing through doing. Indeed, the investigation cycle that is used throughout our units is easily adapted to performance assessments. In the investigation cycle, students are given some problem for which they are to test a series of solutions and recommend one solution. Performance assessments, too, are based on a problem and an activity session during which data is collected that will address the problem. In the investigation cycle, the data gathering and interpretation are followed by critical examination of proposed solutions; performance assessment scoring rubrics can be used to provide feedback directly to students on their own proposed solutions and explanations. The

rubric can be used to orchestrate a discussion of the target concepts and skills, so that students are more aware not only of what the better solutions are, but also the importance and method of using data to explain why a given solution is better. Giving students access to videotaped performance assessment sessions and evaluation rubrics allows students to self-assess their own science practices and collaborative skills; the rubric articulates expectations and keys these expectations to observable actions, so that students have specific actions that they can aim to try or to change.

As we develop structured interviews, we are selecting tasks that have qualities that make them easy for teachers to administer. Namely, these tasks should be easy to perform one-on-one, use little class time, be discreet to avoid distracting the class, have clear guidelines on the types of questions to ask and the extent of follow-up, and explain how to assess responses. We are beginning to consider innovative applications of computer modeling environments, as well as teacher-run interviews.

If we are to replace capstone-type evaluations with embedded assessments, we need to include formats that are familiar to students, teachers, administrators, and other stakeholders. However, the material contained within the familiar format may be more suited to assessment for scaffolding purposes. We are beginning to work with teachers to help them learn to identify aspects of learning that they want to assess for scaffolding purposes, and then to use traditional formats such as written assignments and quizzes as feedback for scaffolding efforts. We are providing teachers with information about resources such as the FMCE (Thornton & Sokoloff, 1998) and the FCI (Hestenes, Wells, & Swackhamer, 1992) so that they can learn to adapt or develop test items that help pinpoint the developmental level of a target concept.

Finally, we seek to find ways to help teachers assess the classroom culture, and especially their own actions and habits within the classroom. We are working to adapt the OPT and the Fidelity of Implementation Report Card for the teachers' use as a self-assessment survey. Helping teachers self-assess is especially important if LBD™ is disseminated more widely, when teacher support and feedback will come from peers and administration instead of directly from the LBD™ development team.

Our goal in assessing and scaffolding the learning in a large scale curriculum implementation has been to address the variability we find across classrooms attempting the

same curriculum units. Our ultimate evaluation results will provide a mosaic of multiple data sources that can help us document and account for that variability. We believe that multiple measures of both the classroom environment and the learning outcomes serve to dovetail these measures. Thus, this approach affords capturing the actual complexity and variability of classroom environments, even those sharing the same teaching and learning goals and the same curriculum materials.

We are encouraged by the results our formative assessments efforts have yielded and we are currently examining the data for more summative results and our preliminary findings look promising. Taken together, we have created a way to capture the complexity in real classrooms attempting serious reform. Our acknowledgement of the variation and unique aspects of each classroom has challenged us to develop the approaches we have reported here. It is our next step to begin the translation of our research efforts into a format that will allow teachers and students to develop ever more autonomy in doing embedded assessments throughout their efforts at project based approaches. We are hopeful that we have established a framework for doing so from what we consider sound theoretical and practical consideration. It is only by being in the classroom and collaborating with our teachers that feel positioned to embark confidently on this next step.

References

Crismond, D., Camp, P.J., & Ryan, M. Design Rules of Thumb—connecting science and design. Symposium talk presented at the *2001 Annual Meeting of the American Education Research Association*, Seattle WA, April 2001 www.cc.gatech.edu/projects/lbd/Conference_Papers/2001_conference_index.html

Fasse, B. & Kolodner, J.L. Evaluating classroom methods using qualitative research methods: Defining and refining the process. In *Proceedings of the Annual Meeting of the International Conference of Learning Sciences*, Ann Arbor MI, June 2000. http://www.umich.edu/~icls/

Fasse, B., Holbrook, J.K. & Gray, J. (1999). *Intermediate Indicators Tool (ITT)* Learning By Design Project document. Georgia Institute of Technology, Atlanta, GA.

Geertz, C. (1983). Thick description: Toward an interpretive theory of culture. In R.M. Emerson (Ed.), *Contemporary Field Research* (pp. 37-59). USA: Waveland Press.

Goetz, J.P. & LeCompte, M.D. (1984). *Ethnography and qualitative design in educational research*. Orlando, FL: Academic Press.

Gray, J., Camp, P.J., Holbrook, J.K., & Fasse, B.B. 2001). Science talk as a way to assess student transfer and learning: Implications for formative assessment. Symposium talk presented at the *2001 Annual Meeting of the American Education Research Association*, Seattle WA, April 2001 www.cc.gatech.edu/projects/lbd/Conference_Papers/2001_conference_index.html

Gray , J. & Holbrook, J. (in preparation). *Student self-assessment in a project based classroom.* Georgia Institute of Technology, Atlanta, GA.

Hestenes, D. Wells, M. and Swackhamer, G. (1992). Force Concept Inventory. *Physics Teacher*, **30**, 159-165.

Holbrook, J.K.; Gray, J; & Fasse, B. (1999). *Observation prompt tool (OPT).* Learning By Design™ project document. Georgia Institute of Technology, Atlanta, GA.

Holbrook, J.K., Fasse, B., Gray, J., & Camp, P. (in preparation). *How the quality of science culture in science class predicts student learning outcomes.* Georgia Institute of Technology, Atlanta, GA

Kolodner, J.L. (1993). *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Lincoln, Y.S. & Guba, E.G. (1985). *Naturalistic inquiry*. CA: Sage.

PALS website, (1999) *Where the Rubber Meets the Road Performance Assessment Task*, developed by the Kentucky Department of Education, available at the website. SRI International, Center for Technology in Learning, http://www.ctl.sri.com/pals/index.html

Spradley, J.P. (1980). *Participant observation.* NY: Holt, Rinehart & Winston.

Thornton, R.K. & Sokolofff, , D.R. (1998) Assessing student learning of Newton's laws, The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, **66**, 338-352.

Wiggins, G.P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass.

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, **20**. 99-149.

Acknowledgments

This research has been supported by the National Science Foundation (ESI-9553583), the McDonnell Foundation, and the BellSouth Foundation. The views expressed are those of the authors.

Appendix 1: DRAFT, do not use without permission

Performance Assessment tasks: Coding for science practice

Additional notes are fine and can be recorded on the coding sheet.
Please note which event segment is being coded for each episode:
planning an experiment; problem set up; experimental manipulation; response to written questions.

In general, the 5 -point likert scale reflects the following quantitative continuum. Details for each item are also included below.

1 = Not at all: no evidence of the quality to be rated
2 = Some evidence that at least one episode or one student exhibits the quality rated
3 = The quality is exhibited half the time
4 = The quality is exhibited for more than half the episodes
5 = The quality completely captures the nature of the episodes

Design an experiment segment:

Within an episode, the context of the group is characterized by:

Negotiations

Not at all	at least one of the members of the group suggests a compromise about some aspect of the procedure	at least one of the members of the group suggests that compromise or debate is needed for at least half the issues that require it	at least two of the members of the group questions several aspect of the procedure and the group makes the needed change	Most decisions are made about procedure by the entire team contributing and decision making is consensual
1	2	3	4	5

Distributed efforts and tasks

Not at all	at least one of the members of the group suggests that others help do the task	at least two of the members of the group suggest that all do some part of the task	at least one of the members of the group suggests and leads the group in dividing and doing the task	More than one of the members of the group enlists the participation of all the team in doing the task
1	2	3	4	5

Level of Understanding of the problem

Not at all	The group thinks the task is to write something down disregarding the "design" aspect	at least two of the members of the group try to work out a method and "run an experiment" with the material available	at least one of the members of the group recognizes that an experiment is to be designed and shares with the other members	More than one of the members of the group enlists the participation of all the team in designing an experiment and that it calls for additional materials
1	2	3	4	5

Use of materials to get to a method

Not at all	At least one member of the group manipulates the material (s) while trying to develop a solution	at least two of the members of the group examine and use the material in a way that might suggest an effort to discover a method	at least two of the team members manipulates the material to explicitly suggest a method	The team explores the material as if messing about to understand what to include in their design/method
1	2	3	4	5

Prior knowledge is defined as students referring to some aspect of the curriculum unit that relates to the current problem; referring to some aspect of a personal experience that seems to relate to the current problem; referring to some aspect of the science concept or method at issue that appears to come from previous exposure to the concept or skill.

Students show evidence of using prior knowledge to solve the problem

Not at all	at least one of the members of the group mentions a prior event or concept that relates to the problem	at least half of the team mentions a prior event or concept that relates to the problem	Several events and concepts are mentioned and applied to the problem	The group routinely recalls events or concepts that assist in their collaborative problem solving
1	2	3	4	5

Prior knowledge seems adequate

Not at all	at least one of the mentions of prior knowledge is followed up on and is useful	At least half the mentions of prior knowledge are appropriate to the problem	More than one member of the group mentions or follows up on events or concepts that are useful	Every mention of prior knowledge is directly applicable to the problem
1	2	3	4	5

Science terms are used in a way that indicates some degree of understanding and can be argued that they are not attributed to the science terms included in the problem description.

Students use science terms to discuss problem solution

Not at all	at least one of the members of the relates the discussion to some science concept	at least half the team relates the discussion to some science concept	Most of the team members use science concepts or terms in such a way that accurate understanding and application are noted	All members of the the team members use science concepts or terms in such a way that accurate understanding and application are noted
1	2	3	4	5

Students use science practice to decide on method/procedures

Not at all	at least one of the members of the group suggest a method to test at least one variable	at least one of the members suggest a method and indicates an understanding of fair testing	at least one of the members suggest a method and indicates an understanding of fair testing and controlling for variables	Most of the team agrees that the method used will fairly test the important variables and their decisions would actually be a reasonable experiment
1	2	3	4	5

The episodes are characterized by group self-checks on procedures

Not at all	at least one of the members of the group questions some aspect of the procedure	at least one of the members of the group questions some aspect of the procedure and the makes the needed change	at least one of the members of the group questions several aspect of the procedure and the group makes the needed change	More than one of the members of the group questions several aspect of the procedure and the group makes the needed change
1	2	3	4	5