# *Learning by Design™ Technical Report*
# *Results of Performance Assessments for 1999-2000 and 2000-2001*

***Jackie Gray and the Learning by Design™ Team***

The AAAS standards (AAAS, 1993) list a variety of science skills and practices to target in middle school education.

- 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)

They list, as well, a set of practices that middle-school students tend to have difficulty with.

- 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 provides extensive practice opportunities and scaffolding for learning this whole set of practices. The data reported herein summarize what we've learned so far about LBD's ability to promote that learning.[1] We analyze the learning of these skills and practices using performance asessessments. Students work as individuals or in groups to address a scientific challenge (e.g., design an experiment, run an experiment, analyze some data). To assess their capabilities in engaging in science practices, we video their activities and the discussions they have with each other and code them for scientific practices the students engage in.

[1] We report on learning of science content in a separate report.

### *Performance Assessments*

Scientific reasoning skills are difficult to measure, but we have had some success in showing the acquisition of these very skills by the students in our LBD™ classrooms. As others have recognized, general strategies emerge out of repeated experience in problems rich in domain content. To fully understand the emergence of scientific reasoning skill, prior knowledge will need to be included in the picture (Zimmerman, 2000). The challenge will be to continue to explore the role of prior knowledge in the acquisition of new knowledge. We have developed a series of problem solving tasks for students in our work to assess their scientific reasoning as well as their collaboration skill. We have gained a great amount of expertise in the past two years in designing 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. Preliminary evidence based on these tasks shows that LBD™'s practices do indeed promote transfer in the subset of the students we have evaluated. Our results also show 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 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.

An example task may help bring this to life. In "Where the Rubber Meets the Road," based on a similarly-named performance-assessment task from the PALS database (http://pals.sri.com/), part 1 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, part 2 has them run an experiment that compares the two types of rubber, and part 3 has them analyze the experimental data and draw conclusions. The science concept being tested is understanding of the force needed to counter sliding friction. A second task asking students to design an experiment to test the velocity of model cars was also used in 2000-2001.

To collect data on performance, 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.

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 the coding scheme developed to assess collaboration and science practice skills during these tasks.).

### *Results for the school year 1999-2000*

When we used our coding scheme to analyze student performance, we found that for an LBD typical achievement levels classroom vs. a similar comparison classroom, there were statistically significant differences in mean scores for the "distributed" and "self checks" measures, and a nonsignificant trend for prior knowledge adequacy. [2] In each case, the LBD means were higher than the comparsion class. For honor students doing LBD vs. a non-LBD honors comparison classroom, there were significant differences for the negotiation, science practice, and self-check

[2] Reliability for the coding scheme ranged from 82-100 percent agreement when two coders independently rated the tapes. For this set of data, a random sample of four tapes for three teachers and three tapes for one teacher (complete set) were used for this analysis. A total of 15 group sessions were coded representing 60 students.

measures with higher LBD means (See Table 1). LBD students were better than comparison students at collaboration, metacognitive awareness of their practices, and ability to remember and use what they had learned previously. Students in LBD classrooms participated in collaboration that can be characterized by negotiation and the distribution of the work. Students in comparison classrooms worked in groups without taking advantage of the unique affordances when work is distributed or solutions negotiated.

This assessment is important for several reasons. First, it tells us that the combination of scaffolding and orchestration that we have developed for LBD is successful in helping students achieving important learning goals. This adds to less formal evidence presented above that design diaries are providing the kinds of scaffolding we predicted. Second, it tells us that we are on the right track in designing the performance tasks and their coding metrics. As these become more concise, we will make them available to teachers and students as scaffolding, showing them the kinds of articulations and practices we expect them to be able to achieve. And third, it provides evidence that these "habits of mind" are being learned and transferred (Kolodner, Gray, & Fasse, submitted).

*Results for the School year 2000-2001*

The coding scheme for our performance assessments having proven to be reliable, it was used to code the data from video taped sessions of LBD and their comparison classes collected at two points during the school year. This year we administered a performance assessment to all our LBD classes and their comparison classes four weeks into the school year. For LBD students, this was at the end of the LBD launcher unit that introduces science and project practices. The comparison classes had been covering similar material during the beginning of the school year during those same four weeks. Additionally, we replicated our 1999-2000 administration for performance assessments after the target unit for both LBD and comparison classes. Table 2 presents the means and standard deviations for a sample set of comparisons. One of our LBD classrooms, representing a typical mix of student ability levels, and a comparison classroom with a similar population of students are presented. For each class, five groups of students with 4 students each (N = 40) were video recorded at the two points during the school year described above. The results indicated that the LBD students were rated

significantly higher then their comparisons after the launcher unit on two (science practice and self-checks) of the seven coding categories. See Table 2.

The results from the analyses of differences after the unit revealed significant differences for all 7 coding catergories where LBD students were rated significantly higher than their comparisons. See Table 2 for means and standard deviations.

### *Current analyses*

We are currently coding the remaining video tapes that represent a large sample of LBD classrooms and their comparisons (LBD typical classrooms for 6 additional teachers and their comparisons for 3 additional teachers representing approximately 60 groups of LBD students and 30 groups of comparison students.) These students are from two separate school districts in the greater Atlanta area; one a suburban community north of Atlanta and one a rural/exurban community south of Atlanta. With the data presented above in Table 2, the total number of students we are evaluating on the coding scheme's measures is approximately 400. We will continue to examine appropriate pair-wise classrooms, matching for student populations (ability levels, SES), school context, teacher expertise, and performance assessment condition. We believe from the trends we reported for 99-00 and the replication of those findings in our first set of analyses for 00-01, that we will be able to report from a large representative sample of students, findings that support our predictions.

### *Discussion*

The data we've collected so far show us that indeed we are having an impact in helping students learn the important science skills and practices that the national standards (AAAS, 1993) list as important. Results show that the orchestration of our Apollo 13 and Vehicles in Motion physical science units, the activities students are working on, and the scaffolding that is provided by teachers and *Design Diaries* have potential for impacting the development of these very important skills. We designed LBD with learning of these skills in mind, and they are routinely practiced in our LBD™ classrooms. Of special interest in our efforts, is the effect of the launcher unit, Apollo 13, on the emerging science practice and metacognitive skills of the LBD students. We are interested in this finding as we continue to examine this effect in the rest

of this year's data. The tremendous gains at post unit for LBD students is also encouraging and our continued analyses may replicate this important effect.

**References**

American Association for the Advancement of Science (AAAS). (1993). Benchmarks for science literacy. Project 2061: Science of all Americans. Washington,DC:Author.

Gray, J., Camp, P., Holbrook, J. and Kolodner, J. (2001*). Science talk as a measure of transfer: Implications for formative assessment.* Paper presented at the Meetings of the American Educational Research Association. Seattle, WA.

Kolodner, J. L., Gray, J., and Fasse, B. (submitted). *Promoting transfer: Rituals and practices in Learning by Design™ classrooms*. Georgia Institute of Technology, College of Computing, Atlanta, GA.

Kolodner, J. L., Crismond, D., Fasse, B. B., Gray, J. T., Holbrook, J., Puntambekar, S., & Ryan, M. (under review). Problem-Based Learning Meets Case-Based Reasoning in the Middle-School Science Classroom: Putting Learning-by-Design™ into Practice. *Journal of the Learning Sciences.*

National Research Council (1996). *National science education standards*. Washington DC:

Puntambekar, S., Kolodner, J.L., (in preparation). From 'Learning to Scaffold' to 'Scaffolding to Learn': A Case for Distributed Scaffolding.

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

Table 1:
***Means and standard deviations for categories from performance assessment coding for LBD students (typical and honors) and Comparison students (typical and honors)***

| **Coding category** | **Means (SD) LBD Typical** | **Means (SD) Comparison Typical** | **Means (SD) LBD Honors** | **Means (SD) Honors Comparison** |
|---|---|---|---|---|
| Negotiations | 2.50 (1.00) | 1.50 (.58) | 4.50 (.58) *** | 2.67 (.58) |
| Distributed Effort/tasks | 3.25 (.50) * | 2.25 (.50) | 4.00 (1.15) | 3.00 (1.00) |
| Prior knowledge | 2.25 (.50) | `1.75 (.50) | 3.75 (1.50) | 3.0 (.00) |
| Prior Knowledge adequate | 2.75 (.96) | 1.50 (.58) | 3.50 (1.00) | 2.67 (1.15) |
| Science terms used | 2.50 (1.29) | 1.75 (.50) | 3.50 (1.00) | 2.67 (1.15) |
| Science practice skills | 2.75 (.96) | 2.25 (.50) | 4.75 (.50) *** | 2.67 (.71) |
| Self-checks | 3.00 (.82) ** | 1.50 (.58) | 4.25 (.50) *** | 2.33 (.58) |

**Significance levels**: * = $p < .03$; ** = $p < .02$; *** = $p < .01$
The means are based on the likert scale: 1 - 5

Appendix 1:

**Performance Assessment tasks: Coding for science practice**

Jackie Gray, Paul Camp, Jennifer Holbrook, Barbara Fasse, and Janet Kolodner

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 |

***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 |