Title: An Efficacy Study of the FOSS/ASK Diagnostic Embedded Classroom Assessment –
Lessons Learned in Implementation

Author(s): Cathy Ringstaff, Michael Timms, & Steven Schneider
**Background / Context:**
WestEd, in collaboration with the University of California, Los Angeles is conducting a four-year study to test the efficacy of the *Assessing Science Knowledge* (ASK) diagnostic formative assessment system, which was developed to be an integral part of the Full Option Science System (FOSS) elementary science modules published by Delta Education, Inc. States and localities are increasingly interested in diagnostic embedded classroom assessment as they look for ways to increase student achievement and to expand their assessment and accountability systems under No Child Left Behind (National Research Council, 2005). Embedding diagnostic classroom assessment practices in instructional materials provides a way for teachers to monitor student learning and provide immediate feedback to students at key junctures. Diagnostic information allows teachers to adjust instruction accordingly (Black & Wiliam, 1998a, 1998b; Black et al., 2002). Curriculum developers, particularly those in science, have begun to develop curricula that employ diagnostic assessment, as science contains complex, sequential sets of knowledge and skills to master. However, the impact of embedded formative assessments on student learning has not yet been rigorously examined in the context of widely implemented commercial science curricula.

**Purpose / Objective / Research Question / Focus of Study:**
The hypothesis being tested in this efficacy study is: Students who are taught in classrooms using the FOSS modules in conjunction with the ASK embedded diagnostic formative assessment system, will exhibit greater improvement in science knowledge than students who are taught with the same curriculum that does not include the ASK embedded assessments. The study, which is in its final year of implementation, utilized a group-randomized design, whereby schools using FOSS were randomly assigned to either a treatment or a control group. Third and fourth grade classrooms in schools assigned to the treatment group are being taught with FOSS modules that include ASK formative assessments and classrooms in schools selected for the control group are being taught with FOSS modules only.

The findings from this study of the ASK diagnostic formative assessment system will not only tell us about the efficacy of this specific system, but will also contribute to a better understanding of the general effects of embedded diagnostic assessment on student learning in commercially available science curricula.

**Setting:**
The study was originally planned to be conducted in just one state, Arizona, in school districts in and around Phoenix and Flagstaff, but challenges in recruitment led to a redesign of the implementation. We extended the study by creating two cohorts, phased one year apart, with schools in cohort 1 in Arizona and additional schools in the states of Washington and Texas. This new implementation design involving three states rather than just one created challenges in recruitment, coordination and outcome measures that we have learned how to overcome.

**Population / Participants / Subjects:**
A total of 103 schools in Arizona, Washington and Texas were recruited to participate in the study. During implementation 24 schools have dropped out of the study for a variety of reasons leaving 79 schools that will complete the study, 40 in the treatment group and 39 in the control group. These 79 schools are in 16 school districts and represents participation of 151 teachers and approximately 3,330 students in third and fourth grade. These large numbers of
participants have required us to design comprehensive administrative procedures to manage implementation of the study and collection of data.

**Intervention / Program / Practice:**

For each cohort, schools in each state (and the teachers within them) were randomly assigned to either treatment (revised program with curriculum-embedded assessments) or control (traditional program) conditions. Treatment teachers participated in two days of summer professional development to orient them to the new curriculum and assessment, follow-up sessions to support the analysis of student work, and a practice year for implementing the curriculum in preparation for the year two test of treatment impact. Control teachers also participated in a similar amount of summer professional development focused on teaching the original curriculum. All teachers in the study implemented two curriculum units, one on Magnetism and Electricity and the second on either Structures of Life or Water.

The FOSS curriculum units used by the treatment group included the Assessing Student Knowledge (ASK) diagnostic formative assessment system. ASK was developed by the Lawrence Hall of Science, University of California at Berkeley, in collaboration with assessment researchers from the University of California at Berkeley Graduate School of Education and SRI International. ASK is comprised of two kinds of assessments: **embedded** and **benchmark**. The embedded assessments are incorporated seamlessly into instruction. They provide continuous information about students’ learning to both teachers and students. The benchmark assessments are summative measures of students’ accumulated knowledge and understanding of science up to that point. These benchmark assessments include a **survey** (pretest) before beginning instruction, **I-Checks** after each major conceptual block in the curriculum (i.e., scientific investigation), and a **posttest** after instruction is completed. I-check benchmark assessments are also used formatively when students self-assess their work. Another component of the ASK approach is for teachers to hold study group meetings with other teachers in their school who are teaching the same unit so that they can jointly review samples of student work and gain insight and expertise in evaluating where students are in learning the key concepts.

The control group used the FOSS curriculum units but without the ASK diagnostic formative assessment system. Both the treatment and control groups attended professional development prior to the start of their teaching for the study. Both groups received training about the general implementation of the FOSS module, but the treatment group received additional training about the ASK system. After the professional development, each teacher taught using his/her respective curriculum for one academic year in what researchers called a ‘practice year.’ The purpose of the practice year was to give each group time to gain experience in use of the curriculum unit and provide a chance to practice the data collection procedures necessary for the study. This practice year proved invaluable in allowing teachers to get used to implementation and for the research team to streamline data collection processes.

**Research Design:**

The study is using a group-randomized design, whereby schools using FOSS are randomly assigned to either a treatment or a control group. To achieve the statistical power needed for the research design, we needed to recruit 70 schools to the study. Schools were the unit of analysis because the use of the ASK system required having two or more teachers at a school so that they could do the study group activities, and it also addressed concerns about contamination across groups. Our initial plan had been to recruit all schools from districts in the
Phoenix, AZ area, but recruitment proved much more challenging than anticipated, so we had to redesign the study into two cohorts, and added two more states to our recruitment efforts in the second year of the study.

Student achievement in science, the key outcome variable in the study, is being measured by two different instruments, (1) a custom developed multiple-choice test aligned to the content of the FOSS modules and (2) pre- and posttests from the ASK assessment system for each FOSS module. Originally, when the study was going to be solely in Arizona schools, we had intended to use the Arizona state test as the first outcome measure, but when we had to extend recruitment to two other states and to both third and fourth grades, this became impossible. Instead, we developed and pilot tested a multiple-choice test that comprised items that matched to the curriculum units used in the study. This approach allows us to compare performances across all schools and has the benefit of providing a reasonable number of items that directly map to the content that is the focus of the study. In the previous approach, we had no control over how many questions on the state test might map to the content covered in this study. The second outcome measure is the pre/posttests (one instrument used twice) from the FOSS curriculum units. These pre/posttests are a part of the ASK system, but the regular FOSS units only used the test as a posttest, so we had to require that the control group students also take it as a pretest.

Fidelity of implementation was measured for both the control and treatment groups. During professional development, teachers participating in the study were made aware of what constituted adequate implementation of the curriculum modules and gathering of data for the study, and they made a commitment to do so. To maintain good fidelity, we employed site coordinators in the states in which the study took place and their responsibilities included monitoring implementation and answering teacher questions. Professional development at the end of the practice year and before the start of the ‘experiment year’ in which the data for the study were collected, reinforced implementation practices. In each lesson, treatment teachers used a document entitled “At a Glance,” a brief description of the important concepts in the lesson. Fidelity was measured through logging of student pretest/posttest and end of year test data and during instruction by weekly completion of a survey in an online teacher log, which asked a range of implementation questions. Failure to complete logs for a week when one was due triggered a reminder email generated automatically by the online system, and continued failure to complete logs prompted an email and/or phone call by the site coordinator.

A second measure of fidelity of implementation was through classroom observations. The classroom observation protocol paralleled the core instructional and assessment components of the curriculum. Observations served as reference for follow-up interviews regarding curriculum implementation. Interviews paralleled the classroom observation components, and provided information on teacher strategies for providing student feedback, making instructional decisions, observing students for patterns, and approaches to analyzing and interpreting student work.

In addition, we measured teacher content knowledge and pedagogical content knowledge about magnetism and electricity, the topic of one of two curriculum units that all study participants implemented. The measure was administered as a pre/post-test, before and after teachers implemented the curriculum twice, in two subsequent years. Three item types corresponded to different aspects of teacher knowledge:

a) content items, as a proxy for teachers’ understanding of science concepts;

b) analysis and interpretation of student work items, as a proxy for teachers’ pedagogical content knowledge; and
c) next-instructional steps items, as a proxy for teachers’ instructional (pedagogical) knowledge.

We are also gathering student demographics and student state assessment performances on mathematics, English language arts, and science standardized assessments to use as covariates in the data analyses.

Data Collection and Analysis:
Data from the pre/posttests and end of year tests are gathered from scantron sheet that students complete, and teachers compile them and send the batches to WestEd for scoring (of the few written response items on the pre/posttest) and scanning. To date we have collected data from Cohort 1, and are awaiting data from Cohort 2 before we can conduct our full analyses. We will be estimating a three-level (students, teachers, and schools) hierarchical linear model (HLM) to estimate intervention impacts.

Findings / Results:
We do not yet have a complete dataset to report findings, but we anticipate that preliminary results will be available by September. In addition to the preliminary results, we can report on the lessons learned about implementing this large study. Our main findings are:

1. **The use of a practice year helped immensely to prepare for a smooth implementation in the experimental year.** Teachers benefitted from being able to put into practice the teaching processes that they learned in the professional development at the start of each of the two units they taught. They also had the opportunity to follow up with questions in a PD session that occurred before the start of the experimental year. Also, the practice year was helpful to the teachers to get used to the data collection in the study.

2. **Appointment of site coordinators was essential to smooth running of the study.** We employed a site coordinator for Arizona and Washington, and two coordinators in Texas where the largest number of participants were. The coordinators helped with logistics for PD in their state, helped to keep teachers moving along through the study, made site visits as necessary, and gave monthly reports to the research team.

3. **Automation of data collection via scantron sheets and online teacher logs helps.** With the large number of teacher and student participants it became essential to automate as much of the data collection as possible, which led to us developing scantron sheets for student response data collection and a web-based teaching log for teachers to complete on a weekly basis as they taught the units.

4. **Use of database to track teacher participation and data submissions** We developed a database to track all aspects of the teacher participation, including their submission of pretests/posttests, participation in PD sessions, stipend payments, etc. This enabled the research team to produce reports on various aspects of the project.

Conclusions:
We have learned a lot about the organization and implementation of this large randomized controlled trial. When we were unable to recruit all the teachers we needed in the first phase of the project and had to create two cohorts it seemed at the time to be disappointing, but on reflection, having a smaller first cohort offered the opportunity to learn how to manage the study. This allowed us to create solid systems that worked to have helped us to ensure strong fidelity of implementation for both the treatment and control groups, maximize the return of data and to prepare for our upcoming data analysis to answer our research questions.
Appendix A. References


Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. London: King’s College London, School of Education. (See also article with the same title, 1998, in Phi Delta Kappan, 80, pp. 139-148.)
