In recent years, the United States Department of Education through law and regulation has placed increased evidence on the use of "scientifically-based research" to demonstrate the effectiveness of educational interventions. In practice this has evolved to a very strong focus on the use of randomized experimental designs in such demonstrations.

While these kinds of experiments can be quite useful, once interventions enter the educational mainstream and become programs, the use of a randomized experiment to evaluate a program is generally inappropriate. These reasons often involve ethical issues. When an intervention is untested, there is no moral obligation to offer the treatment to all participants, because the effectiveness of the treatment is unknown. However, once we have graduated from conducting an experiment to implementing a program, the assumption is that the intervention should be effective, and in such circumstances, it may not be ethical to randomly deny all eligible students an opportunity to participate. In addition, completely randomized experiments often are conducted under artificial conditions that are unlikely to be implemented when the treatment enters broader practice. For example, in an experiment it may be possible to closely monitor the delivery of the intervention or treatment to ensure that the intervention is delivered in the manner intended. While program evaluators also have a strong interest in monitoring the delivery of an intervention, in real-life one will not always be able to monitor delivery closely. As a result, one must devise ways of knowing whether an intervention is successful when delivered in relatively uncontrolled conditions.

In this case, the program in question, a Professional Development Institute offered to mathematics and science teachers through a southeastern university, is one in which evaluating the program through a randomized experiment is clearly not possible. Because teachers volunteer for the program, any post-hoc efforts at a random assignment would be meaningless. Nevertheless, this university has a compelling interest in using a robust design to evaluate the effectiveness of their (and educators) investment of time, treasure, and energy in this project.

For this project, a nationally recognized research, testing, and evaluation facility partnered with the featured university to develop and implement a design to evaluate the effectiveness of a professional development institute for science and mathematics teachers.

Program Description

The professional development institute, PDI, was a year-long experience where mathematics and teachers commit to two weeks of summer involvement (8 days from 8:00-4:00). During this time, teachers experienced standards-based inquiry-based instruction; they learned to critique their own instructional practice; and they worked in teams to develop inquiry-based exemplars that will be implemented during the school year. During the academic year, there were five follow-up experiences where teachers came together to share what was working, trouble shoot challenges, and modified the exemplars that were developed. The fifth session was a Science Technology Engineering, and Mathematics Coalition (STEM) conference designed to explore the interaction of industry and education. The program's goal was to improve student academic achievement in the areas of focus through the implementation of this design. In particular, the expectation of PDI participants was that they would improve their ability to lead the process skills and the specific content areas where exemplars were created and then implemented during the academic year.

The first cohort of teachers started the program in the summer of 2008 and will implement what they have learned during the 2008-2009 school year. New cohorts will be introduced to the program during the 2009-2010 and 2010-2011 school years.

Prior research suggests that interventions like the PDI are effective in helping teachers implement inquiry based instruction. Teacher participating in such programs are able to deliver inquiry based instruction with greater sophistication and accuracy. They are also able to create and implement lessons that score much higher on a the eQUIP protocol, a protocol that measures the quality of inquiry (Author, In Press). Whether a teacher's improved effectiveness at delivering inquiry-based instruction translates to improved student learning is what was measured.

Participants

Twenty-two middle school mathematics and science teachers from several school systems near the university constitute the first cohort for the project. We anticipate a similar number of teachers will participate in the two following cohorts.

Evaluation Design

The goal of the evaluation design was to implement a relatively non-intrusive, carefully targeted evaluation of student learning that would facilitate a robust assessment of the program's effectiveness without unduly burdening teachers and the participating school systems.

Assessment

Because the goal of the PDI is ultimately to improve student achievement, the researchers faced a dilemma that is commonly experienced by those engaged in these kinds of efforts. On the one hand, assessment instruments commonly used in school systems are not necessarily aligned to measure the particular academic goals of the program being implemented. As a result, the use of a state assessment, or traditional standardized test, can introduce noise into the evaluation effort that can make it difficult or impossible to identify whether there was a meaningful academic effect. On the other hand, introducing a well-targeted assessment can be intrusive. Many school systems and teachers already feel students are over-tested, and students may lack any meaningful incentive to offer their best performance on an instrument that is not part of the instructional program.

Because most teachers in our state work in school systems that use Northwest Evaluation Association's Measures of Academic Progress (MAP), a design was implemented that allowed data from this assessment to be extracted and used to evaluate the impact of the PDI on academic learning in the classrooms of the participating teachers. The design of MAP has several strengths in this regard that should be noted:

- MAP is aligned to our state standards in mathematics and science, which minimizes the amount of noise introduced because the assessment may not be aligned to what is expected to be taught. In addition, a single version of MAP is used throughout our state, which permits the use of results from the instrument across districts. Predictive validity between the MAP assessment and state assessments is generally quite high (Author, Kingsbury, Dahlin, Adkins, & Bowe, 2007; Northwest Evaluation Association, 2005a).
- MAP is an adaptive assessment. Because MAP is adaptive no two students take the same form of the assessment. This means, when evaluating group results, that MAP provides a broader, more robust sample of the domain than can be generated from a single fixed form assessment administered to a group of students (Northwest Evaluation Association, 2003). It also makes it easier to study sub-domains of a content area, because a group of

several hundred students will generally provide sufficient numbers of item responses on a large array of items to produce rich, meaningful results.

- MAP uses a Rasch-scaled item pool rather than a scaled test form. For example, there are approximately 5,000 active items in the MAP mathematics item pool. All of these are calibrated to a single cross-grade Rasch-based scale. This calibrated item pool becomes the parent of all tests created from the pool. Thus one can flag portions of a particular test's item pool for analysis and, as long as the items selected adequately cover the scale's range, compare the results of this analysis to other domains and populations.
- MAP uses a cross-grade scale with robust growth norms (Northwest Evaluation Association, 2005b). This facilitates more accurate measurement of student growth across time.

For purposes of this study, the MAP instrument was used as a measure of student achievement. In order to improve the alignment of the test to the particular instructional objectives of the PDI program, content experts from the featured university selected a subset of the items in the state mathematics and science Concepts and Processes item pools that were determined to be aligned to the instructional units implemented by teachers in the program. Group level results (by teacher when count was sufficient, and by program) were rescored and used to report student achievement related to the program achievement. These results could be compared with results on other parts of the MAP assessments, and results achieved by non-participants in the program.

Evaluation Design –

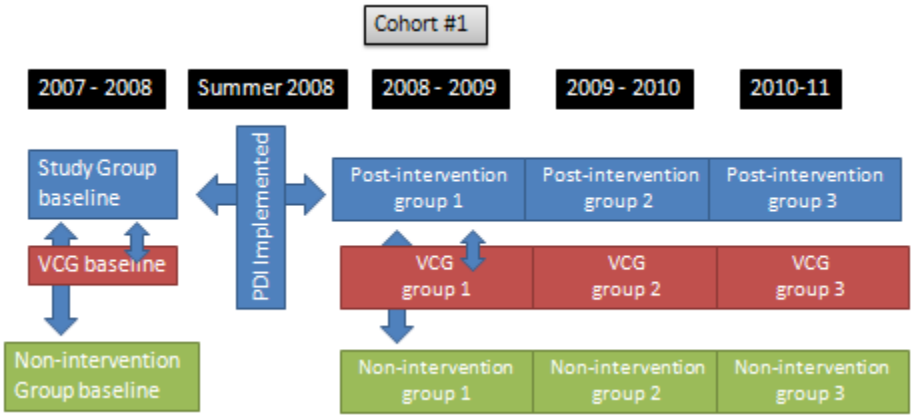
Study Group – For evaluating student achievement gains among the first cohort in the program, the study group will be composed of students who are taught by participating district teachers (large diverse district) who participate in the PDI. As new cohorts of teachers enter the project, they will be followed as additional study groups.

Comparison groups – Two comparison groups will be created for the analysis. One comparison group will consist of students from participating district teachers who did not participate in the PDI. The second will be a Virtual Comparison Group of students matched to the students of the study group teachers. The criteria for creating the Virtual Comparison Group are as follows:

1. Each student in the study group is matched with up to 51 students who serve as virtual comparisons. Students who cannot be matched with at least 21 students are excluded from the analysis.
2. Selected VCG students must have an overall scale score within one point of their study group student.
3. Each VCG student must have been tested within +/- 7 days of their study group student.
4. Each VCG student must come from a school with a Free and Reduced Lunch participation rate that is within 5% of the study group student's school.
5. Each VCG student must come from a school with the same urban/rural designation in the National Center for Educational Statistics Common Core of Data as the study group student's school.
6. Each VCG student must have the same gender and ethnic designation as the study group student.

Figure 1 illustrates the evaluation study design as it would apply to Cohort 1. The design allows for us to measure the achievement growth of each PDI with a baseline group of students taught prior to participation in the program and three subsequent cohorts. The growth within a school year of each of these four groups of students will also be compared to both a Virtual Comparison Group and students of non-participating teachers.

Figure 1 – Design for evaluation of the Clemson Professional Development Initiative



The design controls for several important factors that often confound quasi-experimental study designs:

- Pre-post intervention measurement of the study group introduces controls for any effect introduced by the study group teachers' prior performance. This aspect of the design helps determine whether any gains the program might achieve are the product of improvement in the teacher's instructional effectiveness.
- Pre-post intervention measurement of the students' of non-participating educators helps control for a school or school system effect and the influence that interventions or changes within schools might have on the growth of students during the study period. This helps isolate that any gains reported by the program to the intervention as opposed to other district programs.
- The use of Virtual Comparison Groups introduces a control for effects that might be a product of variance in the student cohorts. The use of matched student groups helps assure that any gains reported by the program are linked to improvements in instruction and not a product of differences among the students taught prior to and after the intervention.
- The continued collection of data for two years after completion of the program makes it possible to determine whether any affect found for the program is persistent. In particular, it permits investigation of whether there is a "J-curve" effect associated with this kind of intervention. The J-curve phenomenon suggests that as a new reform is implemented that a lag, and even a slight drop, can be expected until the teacher becomes comfortable with the changes (Erb & Stevenson, 1999). If the reform is effective, student outcomes will improve in the long run, provided that sufficient time is allowed to overcome the J-curve effect (Yore, Anderson, & Shymansky, 2005).

The employment of Virtual Comparison Groups in evaluation studies can be a realistic, robust, non-intrusive option for option for conducting investigations around the effectiveness of

school or university programs that are aimed at improving student learning. The results of this particular project will be shared. They provide helpful formative and summative evaluative information that will be useful in improving the evaluation of professional development programs such as the one on-lined in this paper. The details of the findings, lessons learned, and future refinements will be shared.

References:

Author. (In Press). K-12 science and mathematics teachers' beliefs about and use of inquiry in the classroom *International Journal of Science and Mathematics Education*.

Author, Kingsbury, G. G., Dahlin, M., Adkins, D., & Bowe, B. (2007). *Alternate methodologies for estimating state standards on a widely-used computerized adaptive test*. Paper presented at the National Council on Measurement in Education.

Erb, T. O., & Stevenson, C. (1999). Middle school reforms throw a J-Curve: Don't strike out. *Middle School Journal*, 45-47.

Northwest Evaluation Association. (2003). *Technical manual*. Lake Oswego, OR: Author.

Northwest Evaluation Association. (2005a). *NWEA Reliability and Validity Estimates: Achievement Level Tests and Measures of Academic Progress*. Lake Oswego, OR: Author.

Northwest Evaluation Association. (2005b). *RIT scale norms*. Lake Oswego, OR: Author.

Yore, L., Anderson, J., & Shymansky, J. (2005). Sensing the impact of elementary school science reform: A study of stakeholder perceptions of implementation, constructivist strategies, and school-home collaboration. *Journal of Science Teacher Education, 16*(1), 65-88.