A Science Lesson Plan Analysis Instrument for Formative and Summative Program Evaluation of a Teacher Education Program

Christina L. Jacobs, Penn Science Teacher Institute, University of Pennsylvania
Sonya N. Martin, Queens College, City University of New York
Tracey C. Otieno, Penn Science Teacher Institute, University of Pennsylvania

Abstract

In evaluating the success of teacher development programs, measures of teaching practice that are valid, reliable, and scalable are needed. We have developed, validated, and piloted the Science Lesson Plan Analysis Instrument (SLPAI) for quantitative evaluation of teacher-generated multi-day lesson plans. The SLPAI was developed to complement traditional evaluation tools, such as teacher surveys and direct observational protocols, to enable us to capture the extent to which a teacher development program successfully addressed its goals of increasing teacher content and pedagogical knowledge and impacting teaching practice. This paper presents the development and validation of the SLPAI, and demonstrates its use in a pilot study examining teacher change as a result of program instruction. The SLPAI was utilized as a formative assessment, providing baseline information about the teaching practices of incoming program cohorts in order to tailor both pedagogical and content instruction appropriately. The SLPAI was also used to track and describe changes in teaching practice and pedagogical knowledge of teacher participants over time, and thereby provide summative evidence of program effectiveness. We report on the responses of several program instructors to these results, including revisions made to instructional design of their courses.

Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, April 15-18, 2007, New Orleans, Louisiana.

Introduction and Purpose

Recent years have seen a surge of public interest in the training and subject-matter competency of secondary math and science teachers; this interest has resulted in congressional hearings and high-profile publications such as *Rising Above The Gathering Storm: Energizing and Employing America for a Brighter Economic Future* (National Research Council, 2006). Politicians, business leaders, and institutes of higher education have taken notice of the crucial link between the training of qualified secondary teachers and the production of a well-prepared college student population and workforce. The Math Science Partnership (MSP) granting program at the National Science Foundation (NSF) is a direct outgrowth of this movement. Universities, school districts and other partners have been awarded large sums of money based on the assumption that the activities of such partnerships will result in a higher quality teaching force, which will in turn result in better prepared students.

However, in order to test these assumptions and validate the expenditures of the MSP program, MSP researchers and evaluators must detect and measure change in teachers and their students *and* be able to attribute those changes to the activities of the MSP. For example, the Penn Science Teacher Institute (Penn STI), the institutional context for this study, aims to “increase the content knowledge of science teachers, and change the teaching and learning methodologies used in secondary science classrooms to research-based promising pedagogical practices” (Penn STI web site, para. 2). These two proximal goals are hypothesized to support "the overriding goal [which] is to increase the interest in and learning of science by all students” (Penn STI web site, para. 2). Penn Institute has been touted as a model program for other universities who wish to strengthen their contribution to teacher preparation (National Research Council, 2006). However, at this early stage of program implementation, we cannot offer the MSP or higher education communities a great deal in the way of evidence that our approach is better able to support students' learning of, and interest in, science than other established teacher education programs in many places around the country. Each individual MSP grantee, as well as the research projects that are evaluating the MSP program as a whole, must continue to collect meaningful data and analyze their findings in order to test and hopefully support the basic assumptions of the MSP funding program.

A problem we, and presumably other MSP funded programs face, is *how* to evaluate our progress towards the goals described above. A commonly used approach relies on standardized and usually decontextualized tasks such as surveys, tests and other assessments for teacher and student beliefs and knowledge, often used in a pre/post comparison fashion. This approach can produce large data sets in order to detect and measure change due to an intervention; however, such tasks usually occur outside of the normal practices of a classroom teacher and his or her students and therefore may not be reliable predictors of a teacher’s or student’s actions in context. On the other hand, truly authentic assessment or evaluation depends so heavily on context that any sense of control of extraneous factors becomes impossible. Given these seemingly incompatible goals of inference of causality and authenticity, program evaluation appears at first glance to be an impossible task.

Obviously, though, program evaluation is alive and well. “Mixed methods” research and evaluation has gained attention as a way to address these issues by using a combination of evaluation techniques and instruments. For example, the evaluation plan for the Penn STI includes yearly student and teacher surveys of practices and attitudes, student achievement tests, annual classroom observation and interviews of a subset of teachers using both a quantitative evaluation protocol and qualitative field notes. Upon examining this set of tools, however, we found a weakness in our ability to evaluate the quality of teaching. While classroom observations are useful in this regard, they cannot be implemented at a large enough scale to get a clear idea of even a subset of teachers' classroom practices. Another, more scalable and broader lens with which to view teaching behaviors and the beliefs they evince was needed. This need led to the development of the Science Lesson Plan Analysis Instrument (SLPAI), which provides several benefits as an addition to the comprehensive evaluation plan. It allows for evaluation of longer 'chunks' of planned instruction, allowing insight into the teachers' decisions about sequence of and relationships between activities and topics as well as their assessment strategies, neither of which are commonly evident when observing a single class period. We do not claim that lesson plan analysis is a suitable replacement for either in-depth site visits and classroom observation, or for large-scale administration of pre/post surveys and tests. In an ideal situation, all of these protocols would be used in combination, in a manner responsive to the size and staffing of the project, in order to provide the most reliable and useful information.

This paper describes the SLPAI and its use in evaluating the work of Penn STI teacher participants before and during their involvement in the program. We discuss how results from this instrument fit with information gathered using other established protocols, as well as how it contributes unique and useful data about our participants and our program.

Background

Previous investigations of changing teacher practice as a result of professional development have utilized several techniques to measure teacher practice, either directly or indirectly. Self-report questionnaires have the benefit of ease of administration and the possibility of large sample size, aiding statistical analysis. Self-report questionnaires, triangulated with data from student surveys, were successfully utilized to monitor the effects of a statewide systemic initiative on teacher practices and student attitudes (Scantlebury, Boone, Butler-Kahle & Fraser, 2001). However, research clearly indicates that *how* a teacher understands and implements a reform, not merely its presence in his or her practice, influences the effectiveness of that reform in the classroom (Brown & Campione, 1996). One limitation of using survey data alone, therefore, is that the researcher cannot distinguish between high and low implementation quality of a strategy based only on a teacher's assertion that the strategy is utilized. For example, one survey item used in the 2001 study cited above asked teachers to rate the extent to which their students "use data to justify responses to questions" on a scale from "Almost Never" to "Very Often". Assuming the item is understood by the teacher in the intended way, such "justification" of students' conclusions could reflect greatly varying expectations of students. Triangulation with a similar question for students, "In this class my teacher asks me to give reasons for my answers," is of little use in this case, because the student item does not refer to the use of data and could be interpreted differently by students, depending on their past experiences with writing explanations. From the students' point of view, a "reason" could simply be, "because of gravity," or "because I learned it last year."

A second limitation of self-reported survey data is that a teacher with a higher level of awareness of teaching reforms, and the philosophical and sociological arguments that underlie them, would be expected to have a better understanding of the survey items' intent. This would result in more accurate self-reported data; such teachers would be less likely to rate themselves positively (or negatively) for the wrong reasons. Teachers with little or no background in educational theory, on the other hand, would be vulnerable to misreporting. Using the previous example survey item as an example, a teacher may not be aware of how to structure student inquiries to allow for students to support their conclusions effectively using evidence. Instead, they might have students treat textbook information or their lecture notes as "evidence." Since the way a teacher interprets this and similar items depends on their own level of knowledge about the topic and/or effective strategies for teaching the topic, survey results alone may not give an accurate picture of the extent of reform-oriented teaching in such a teacher's classroom.

Direct observation by a trained evaluator using an instrument, such as the *Reformed Teaching Observation Protocol* (RTOP; Piburn et al., 2002; Sawada et al., 2002) or the *Approaches to Teaching Inventory* (Trigwell & Prosser, 2004), provides an excellent solution to the issues discussed above. The primary drawback of this approach is that classroom evaluation of teachers is resource intense and therefore not highly scalable for evaluative purposes in large programs. In our specific case, our program will have at maximum 120 teacher-participants (6 cohorts of 20 each) under evaluation during an academic year. Given current staffing levels and the geographic area over which our teachers are spread, we are only able to observe and use RTOP to evaluate about one third of the participants from each cohort. Furthermore, since each teacher selected for observation can only be visited once per year, observation data cannot provide a truly representative picture of teaching practices for any individual teacher or the cohort as a whole. For example, a hypothetical lesson sequence might include a day of mainly lecture as new topics are introduced, followed by a laboratory experience on the second day, and a discussion of the lab results on day three. Depending on which day of the sequence was observed, a teacher could receive very different ratings of classroom practice. We have found this to be true in our own program as well; one Institute instructor was observed teaching the same class to the same students on three different days during a semester, and the resulting RTOP scores were quite varied (63, 40, and 77 out of 100), due to the daily activity structure. What is needed to address these problems is an evaluation method with a closer link to actual teaching practices than survey results, but that allows evaluation of all teachers in the program, and provides a longer time frame than a single day of instruction. We believe that the SLPAI fulfills these needs. Using the SLPAI in concert with teacher and student questionnaires for all participants, in addition to RTOP for a subset, will therefore allow us to develop a more complete and accurate picture of the effects our programs have in the classrooms of our teacher-participants.

There are a few examples of lesson plan evaluation present in education literature. The *Science Lesson Plan Rating Instrument* (Hacker & Sova, 1998) focused on procedural aspects of lesson planning such as identification of resources used, timing estimates for activities, and inclusion of lesson objectives on the course syllabus. Of the thirty-four equally weighted items on this instrument, fifteen of them address substantive issues about *how* science is taught. We drew on these categories in developing the

SLPAI; for example, “Have key questions for students been identified?” was folded into our items dealing with classroom discourse and goal orientation, and “Are the selected activities at the correct level of difficulty for the class?” is consonant with our “Content presentation” item.

Regardless of the measurement instrument used, its alignment with the reform agenda or teacher education curriculum being studied is vital. To this end, we have utilized contemporary educational research and reform documents that underpin the mission of the Penn STI to inform development of the SLPAI. We were influenced by the description of learner-, knowledge- and assessment-centered learning in *How People Learn* (Bransford, Brown, & Cocking, 1999). The instrument is aligned with the Science Teaching Standards (A-E) from the *National Science Education Standards* (National Research Council, 1996). Brown and Campione’s review (1996) of the critical features of powerful learning environments also influenced development of the SLPAI, especially many of the items in the “Sociocultural and Affective Issues” category. The SLPAI’s approach to curriculum and lesson design was guided by *Understanding by Design* (Wiggins & McTighe, 2001). Finally, the SLPAI “Nature of science” item was developed out of the extensive literature on teachers’ beliefs and instructional practices around the nature of science (Brickhouse, 1990; Chinn & Malhotra, 2002; Crowther, Lederman & Lederman, 2005). Many of these sources also informed the development of the other instruments utilized in our evaluation, making the SLPAI a theoretical complement to these methods.

Methodology

Setting and Participants

Study participants were in-service teachers enrolled in a Math-Science Partnership (MSP) Institute program funded by the NSF. The Institute is comprised of two Masters degree-granting programs: one for high school chemistry teachers (HSChem) and the other for middle-grades science teachers (MSSci). Each program spans three summers and the two intervening academic years, and requires completion of eight specially designed science content courses and two science education courses over this 26 month period.

All teacher-participants from the 2005 incoming Institute cohorts who were enrolled from partner schools and complied with lesson plan submission guidelines were included in this study. Additionally, data from participants selected for classroom observation from the 2006 incoming cohorts were utilized in the RTOP validation portion of the study only. The analyses presented here are therefore based on 20 MSSci and 8 HSChem teacher-participants in the 2005 cohort, and 7 MSSci and 7 HSChem teacher-participants in the 2006 cohort. Only teachers from partnered school districts were included in this study, due to their availability for classroom visitation as well as program staffing limitations that precluded data collection beyond the scope required for our external evaluation. Some participants who were otherwise qualified were omitted from certain analyses to allow paired statistical testing.

These participants taught in three states and many different districts; we evaluated science lesson plans ranging from grades 5-12. While a large proportion of the teacher-participants worked in urban schools, suburban and rural schools were also represented in the sample. The teachers had a wide range of experience levels (two to twenty-five years of teaching). They were also diverse with respect to their prior science backgrounds, having taken from one to fourteen post-secondary science courses prior to enrolling in the MSSci or HSChem programs.

The authors were employed by the project, either at the time of the data collection (Christina and Tracey) or previously (Sonya), as internal evaluators. As such, they were involved in gathering information for the dual purposes of informing program development and implementation in a formative sense and formal summative program evaluation. Lesson plan analysis and classroom observations were performed by Christina and Tracey.

Instrument Development and Testing

The SLPAI was adapted from a general lesson plan review protocol provided by the Institute's external evaluation team. Based on results from pilot implementations, we refined the wording of several items and added additional items specifically dealing with science instruction, and also significantly modified the scoring mechanism to avoid subjective holistic evaluation and therefore improve inter-rater reliability. Instrument development was an iterative process, in which reviews of lesson plans from teachers not involved in this study were used to refine and specify the rubric wording, organization and scoring protocol.

The SLPAI consists of four major sub-scales: Alignment with Endorsed Practices (AEP), Lesson Design and Implementation – Cognitive and Metacognitive Issues (CMI), Lesson Design and Implementation – Sociocultural and Affective Issues (SCAI), and

Table 1
SLPAI Items by Category with Scoring Weights

Sub-scale Item	Weight
Alignment with Endorsed Practices (AEP)	
Alignment with standards	1
Awareness of science education research	1
Lesson Design and Implementation – Cognitive and Metacognitive Issues (CMI)	
Goal orientation	3
Content accuracy	2
Content presentation	3
Pre-assessment	2
Meaningful application	2
Student reflection	2
Assessment	3
Lesson Design and Implementation – Sociocultural and Affective Issues (SCAI)	
Equity	1
Student Engagement	2
Appropriate use of technology	1
Adaptability	1
Classroom discourse – fostering a community of learners	3
Variety and innovation	2
Portrayal and Use of the Practices of Science (PUPS)	
Hands-on exploration	2
Nature of science	3
Student practitioners of scientific inquiry	3
Analytical skills	3
Error analysis	1

Portrayal and Uses of the Practices of Science (PUPS). The full list of item titles by category is provided in Table 1. A sample item with rating descriptors is shown in Figure 1. For each item, teachers could be rated as Exemplary (2 points), Making Progress (1 point), or Needs Improvement (0 points), or as intermediate between two of these categories. Raw scores were multiplied by item weight coefficients (values ranged from 1-3), which were determined by evaluators according to the goals of the Institute, and were meant to provide flexibility in adapting the SLPAI to other contexts. The weighted item scores were added, and the point total was normalized to give a score out of 100, so that non-applicable items could be excluded when appropriate without negatively affecting the overall score.

Item	**Exemplary**	**Making Progress**	**Needs Improvement**
Student practitioners of scientific inquiry (Weight = 3)	Students are consistently engaged first-hand in learning content through inquiry or *doing*, rather than being told "answers"; inquiry process skills are taught in context.	Some effort at engaging students in *doing* science is evident, with an emphasis on *telling* students science. OR Inquiry is taught out of context as a separate content area rather than as a set of process skills to be applied.	Students learn science exclusively by *being told* accepted scientific knowledge without discussion of how the knowledge was developed.

Figure 1. Example SLPAI item.

The reliability of the SLPAI was examined using independent double-scoring of 25% of the lesson plans (10 of 40) by the co-developers of the instrument. The average inter-rater agreement in this test was 96%.

Subsequent to the completion of the studies presented in this paper, a researcher who had not been involved in the instrument's development was trained to use the SLPAI in order to more convincingly verify instrument reliability. Using a new set of lesson plans, one of the instrument developers achieved 89% inter-rater agreement with the new researcher on 30% (8 of 27) of a new set of plans, submitted by teacher-participants who were not subjects of this study.

Baseline Diagnostic Pilot Study

Prior to beginning their Institute programs, teacher-participants from the 2005 cohorts were asked to submit a sample of a previously enacted unit lesson plan of approximately 5 days in length that they were likely to continue teaching in future years. In addition to a description of the daily lesson activities, they were asked to include copies of assignments, handouts, laboratories and assessments, with examples of graded student work if possible. In this way, planning and some aspects of enactment of the lesson unit could be measured, either directly or indirectly. Our aims were to determine the content and pedagogical areas of strength and weakness of the cohort teacher-participants according to their average total and item scores, and to provide a baseline measure of teaching practice in order to detect change over the course of their studies. All "Baseline" lesson plans that were submitted with sufficient information for review (17 of 21 MSSci plans and 8 of 14 HSChem plans) were evaluated using the SLPAI. The remaining participants either did not submit baseline lesson plans, or submitted materials without enough detail for review using the SLPAI; for example, several teachers simply photocopied their district curriculum guide without indicating which of the suggested activities or assessments were actually used. These teachers were omitted from the analysis.

Teacher Change Pilot Study

MSSci teacher-participants in the 2005 cohort also submitted lesson plans to fulfill a science education course assignment near the end of their first full year in the program. At this point in time, the participants had completed one course each in physics and mathematics, and nearly finished courses in chemistry and science education. The science education course instructor provided the researchers with copies of these teacher-participants "Year 1" plans, which covered 2-3 days of instruction, and they were scored using the SLPAI. Total and sub-score averages were compared to the Baseline scores for significant change using *t*-tests (N =17 Baseline, N = 18 Year 1). Additional item-level analysis using repeated measurement ANOVA was also carried out to find specific areas of change; for this analysis, only teachers for whom we were supplied both Baseline and Year 1 plans were included (N = 15).

Because the course assignment that was used as the Year 1 data source did not require that the lesson had been implemented in a classroom or entail submission of graded student work, analysis of these plans could not provide any information about lesson enactment. Furthermore, these lesson plans were submitted for a different purpose (graded course requirement versus un-graded baseline establishment). For these reasons as well as the difference in length of the plans, we treated the differences between Baseline and Year 1 lesson plans conservatively when attempting to draw conclusions about teacher change.

Other Data Sources

Validity of the SLPAI was examined by triangulation of the results with other measures of teaching practice, including direct observation of a subset of teachers using the *Standards-Based Teaching Practices Questionnaire* (SBTPQ, Scantlebury et al., 2001), and RTOP (Sawada et al., 2002). Validation against the SBTPQ was conducted by comparing cohort-level conclusions generated from the 2005 baseline administration of that survey to conclusions reached using the SLPAI. Validation against the RTOP was conducted at the level of the individual teacher, by testing for correlation between SLPAI scores and RTOP scores on related items.

Results and Discussion

Validation of the SLPAI

Teacher-participants in the 2005 MSSci and HSChem programs that were evaluated using the SLPAI were also administered the previously validated SBTPQ prior to their participation in the program. The results from the independent SBTPQ analysis was compared with the SLPAI data, and similar but not entirely overlapping conclusions were reached. We present here the comparison between SBTPQ responses and SLPAI Baseline data for both HSChem and MSSci teachers as one means for instrument validation.

Using the SBTPQ, external evaluators found that MSSci teachers reported significantly more frequent use of standards-based teaching practices than the HSChem teachers (Table 2), both in terms of what they do in class and what their students do. MSSci teachers were significantly more likely than HSChem teachers to report arranging seating to facilitate student discussion, using open-ended questions, requiring students to supply evidence to support their claims, encouraging students to consider alternative explanations, and using non-traditional or authentic assessments. MSSci teachers also reported that their students were more likely than HSChem teachers' students to design activities to test their own ideas and to talk with one another to promote learning.

Table 2
SBTPQ Items with Significant Baseline Differences for MSSci and HSChem Teachers

	MSSci (N = 23)		**HSChem** (N = 18)		
SBTPQ Item	Mean	SD	Mean	SD	**t-value**
I arrange seating to facilitate discussion.	4.17	1.07	3.06	1.26	9.43**
I use open-ended questions.	4.26	0.75	3.61	0.85	6.73*
I require that my students supply evidence to support their claims.	4.35	0.65	3.83	0.86	4.80*
I encourage my students to consider alternative explanations.	3.73	0.88	3.44	1.15	5.52*
My students design activities to test their own ideas.	2.82	0.80	2.11	0.90	6.95*
My students talk with one another to promote learning.	4.14	0.77	3.56	0.92	4.69*

Note. The items were rated on a 1-5 point Likert scale, where 1 = "Almost Never" and 5 = "Very Often". From *Evaluation of University of Pennsylvania Science Teacher Institute – 2005-2006* (p. 15), by J. B. Kahle and K. C. Scantlebury, 2006, Oxford, OH: Miami University Evaluation & Assessment Center for Mathematics and Science Education. Copyright 2006, Miami University. Adapted with permission.
* $p < 0.05$, ** $p < 0.01$

The conclusions drawn from the SQTPQ measure were also supported by the baseline SLPAI data. Four SLPAI items were identified to address the same teaching practices listed in Table 2: Student engagement, Classroom discourse, Student practitioners of scientific inquiry, and Analytical skills. The Baseline results for MSSci and HSChem teachers in these four items were analyzed using *t*-tests to detect significant score differences between cohorts (Table 3). We found that MSSci teachers scored significantly higher than HSChem teachers on all items. The largest average score difference between MSSci and HSChem teachers was in the promotion of active student engagement ($p < 0.01$). These results indicate that the newly developed SLPAI has diagnostic overlap with the well-studied SBTPQ, thereby giving us confidence in the validity of the four SPLAI items that were compared.

Table 3
Baseline Results for MSSci and HSChem Teachers on SLPAI Items Related to the SBTPQ Items in Table 2

SLPAI item & description	**MSSci** (N = 17)		**HSChem** (N = 8)		***t*-value**
	Mean	SD	Mean	SD	
Student engagement – requires active participation of students in their own learning	1.51	0.49	1.00	0.46	2.81**
Classroom discourse – lesson is structured to require and promote sense-making discussion among students	1.29	0.58	0.84	0.67	2.08*
Student practitioners of scientific inquiry – inquiry skills are taught in context	1.06	0.77	0.50	0.60	2.15*
Analytical skills – students are supported in drawing or refuting conclusions based on evidence	1.07	0.65	0.97	0.85	0.92

Note. Lesson plans were evaluated on a 0-2 point scale.
* $p < 0.05$, ** $p < 0.01$

A second validation test of the SLPAI was conducted using RTOP data. Teacher-participants who also submitted lesson plans for review were observed either via video footage from their baseline portfolio (7 MSSci and 7 HSChem teachers from the 2006 cohort) or in person during their first academic year in the program (8 MSSci teachers from the 2005 cohort). It is important to note that the lessons that were observed were for the most part not taken from the same unit covered by their evaluated lesson plans, but were generated during the same school year as the directly observed lesson. Participants' RTOP scores on items that had previously been determined by the researchers to conceptually align with SLPAI categories were tested for correlation with the teacher-participants' SLPAI scores. The results of this analysis are given in Table 4.

The following SLPAI items exhibited small ($.15 < r < .30$) positive correlations with directly related RTOP items, as expected: Pre-assessment, Meaningful application, Student reflection, Student engagement, and Classroom discourse. As an item that spans many aspects of the classroom environment, "Classroom discourse" was tested for correlation to a number of RTOP items; interestingly, it did not show any correlation with items such as number 20, "There was a climate of respect for what others had to say," which describe teaching practices that could be difficult or impossible to capture in a written lesson plan. Moderate ($0.30 < r < 0.50$) to large ($r > 0.50$) positive correlations were detected between the following SLPAI items and their corresponding RTOP items: "Hands-on exploration", "Nature of science", and "Student inquiry". We were pleased to note that these results were clustered in areas directly related to science teaching, and not surprised that more significant correlations were detected between items clearly observable in both written lesson plan and direct observation formats, such as student use of manipulatives to represent phenomena.

Our comparison of SLPAI and RTOP data for validation purposes also provided some unexpected results. It was assumed that the SLPAI item "Goal orientation", which sought explicit, comprehensive and clear learning goals, would correlate with RTOP items six and seven: "The lesson involved fundamental concepts of the subject," and "The lesson promoted strongly coherent conceptual understanding." However, these items showed a moderate *negative* correlation to SLPAI Goal orientation. Likewise, the expected correlation between between the "Content accuracy" and "Content

Table 4

Correlations Between SLPAI Items and Related RTOP Items (N = 22)

	RTOP item																							
SLPAI item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**24**	**25**
Goal orientation	—	—	—	—	—	-.29	-.20	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Content accuracy	—	—	—	—	—	.07	—	.16	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Content presentation	—	—	—	—	—	—	.06	—	-.21	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Pre-assessment	.27	—	—	—	.12	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	.12
Meaningful application	—	—	—	—	—	—	—	—	—	.23	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Student reflection	—	—	—	—	—	—	—	—	—	—	—	—	—	.29	—	—	—	—	—	—	—	—	—	—
Student engagement	—	—	—	—	—	—	—	—	—	—	—	—	.07	—	—	—	—	—	—	—	.16	.28	—	—
Classroom discourse	—	.29	—	—	—	—	—	—	—	—	—	—	—	—	.24	.10	-.06	.29	.11	-.09	—	—	.11	-.05
Hands-on exploration	—	—	—	—	—	—	—	—	—	—	.34	—	—	—	—	—	—	—	—	—	—	—	—	—
Nature of science	—	—	—	.40	—	—	—	—	—	—	—	—	—	—	.00	—	—	—	—	—	—	—	—	—
Student inquiry	—	—	.44	—	—	—	—	—	—	—	—	.73	—	—	—	—	—	—	—	—	—	.30	—	—
Analytical skills	—	—	—	—	—	—	—	—	—	—	—	—	—	—	-.02	—	—	—	—	—	—	.02	—	—
Error analysis [a]	—	—	—	—	—	—	—	—	—	—	—	—	.08	—	—	—	—	—	—	—	—	—	—	—

Note. RTOP item descriptions can be found in Piburn et al. (2002). Pearson correlation coefficients were calculated for pairs of items determined to be conceptually linked by the authors. RTOP item 23 is omitted from the table because it was not predicted to correlate to any SLPAI item.

[a] Lesson plans involving topics that did not allow for analysis of experimental error were omitted ($N = 15$).

presentation" SLPAI items and RTOP items six through nine (all related to content knowledge and presentation of the teacher) did not materialize, even though the relationship between the items' intentions is clear. This disconnect likely is related to the subtle but insurmountable distinction between the lesson plan and the lesson enactment; use of these instruments in concert may provide a lens with which to address this perennial concern regarding evaluation of teacher practice. Finally, the SLPAI items "Analytical skills" and "Error analysis" were not correlated to their counterpart RTOP items. These mismatches in the data could be due to the generous interpretation of what counts as analytical skills used by the SLPAI in the first case, and the extremely low average score (0.3 out of 2) in the second.

Baseline Diagnostic Pilot Study

Baseline lesson plans from two cohorts of teachers were analyzed using the SLPAI in order to provide information on the strengths and weaknesses of the incoming teachers' knowledge and practices. This analysis was intended to provide program faculty with useful knowledge about the skills and practices of their students in order to gear their instruction to be most effective. Table 5 presents the results for items with either low or high cohort averages (less than 1.0 or greater than 1.5, respectively) and with scoring weight greater than one.

TABLE 5
Item Analysis of Baseline SLPAI Results by Program

SLPAI item & description	**MSSci** (N = 17)		**HSChem** (N = 8)	
	Mean	SD	Mean	SD
Content accuracy	1.66	0.58	1.69	0.53
Content presentation – level of detail and abstraction, sequencing, examples	1.54	0.49	1.53	0.67
Nature of science – tentative nature of knowledge based on changing evidence, social process involving argumentation	0.68	0.52	0.00	0.00
Student engagement – requires active participation of students in their own learning	1.51	0.49	1.00	0.46
Pre-assessment – teacher solicits student ideas in order to plan instruction	0.32	0.56	N/A [a]	N/A
Classroom discourse – lesson is structured to require and promote sense-making discussion among students	1.29	0.58	0.84	0.67
Variety – Teacher innovation or creativity keeps teacher and students engaged	1.47	0.57	0.97	0.66
Student practitioners of scientific inquiry – inquiry skills are taught in context	1.06	0.77	0.50	0.60
Analytical skills – students are supported in drawing or refuting conclusions based on evidence	1.07	0.65	0.97	0.85
Student reflection – students reflect on and summarize their understanding	1.19	0.65	0.71	0.55

Note. SLPAI items were evaluated on a 0-2 point scale.
[a] The "Pre-assessment" item was added to the SLPAI during a later round of revisions, after HSChem Baseline lessons had been evaluated.

From these data, we see that both cohorts were very strong in the areas of content accuracy and content presentation in a self-selected topic area. However, teachers in neither program showed evidence of attention to representing the nature of science, as demonstrated by their low average scores on this item. In addition, MSSci teachers also performed well on the item dealing with student engagement, but poorly on the pre-assessment item. HSChem teachers' lesson plans were below an average score of 1.0 in the areas of classroom discourse, variety, student inquiry, analytical skills and student reflection.

These data suggest that teachers enter our program with established practical knowledge and experience that can be used as a foundation for further growth. We found both groups of teachers to utilize fairly accurate science information and present science topics in a relatively clear and appropriate manner, at least in the areas of science they chose to present in their Baseline lesson plans. We believe that the intensive science coursework provided in the Institute will enable teachers to expand their comfort level with science, improve the accuracy of their teaching diverse topics, bring to the classroom topics they previously avoided, and gain skills and attitudes that favor life-long learning required of science teachers in a technologically-oriented society.

We also found that in several areas mentioned above, our teachers' lesson plans do not include evidence of the social constructivist teaching practices and beliefs espoused by the Penn STI. Teachers may be unaware of or inexperienced in implementing such practices, or their own beliefs about teaching and learning may not be congruent with those of the Institute. These results and possible interpretations point to the need for the Institute programs to address educational theory and the link between the content knowledge learned in the program and that which is utilized in the secondary classroom.

MSSci Teacher Change Pilot Study

The total score distributions of the MSSci Baseline and Year 1 lesson plans are represented at a coarse level of analysis in Figure 2. The number of participants in each score range is depicted for Baseline and Year 1 lesson plans, showing the change in score distribution over time. The score distribution shifted upwards after one year of instruction, showing increases in both the lowest and highest scores and an increase in the mean total score (see Table 5). Broken down by rubric category, significant score increases were seen the AEP and CMI categories using unpaired t-tests, and smaller increases were measured in the SCAI category and the total score (Table 6). No change was seen in the PUPS category although it had the lowest Baseline category average. Note that this comparison is between non-paired samples (some teacher participants were included in only one analysis), and that the Year 1 plans were submitted for a course grade; these differences could account for some of the change seen at this level of

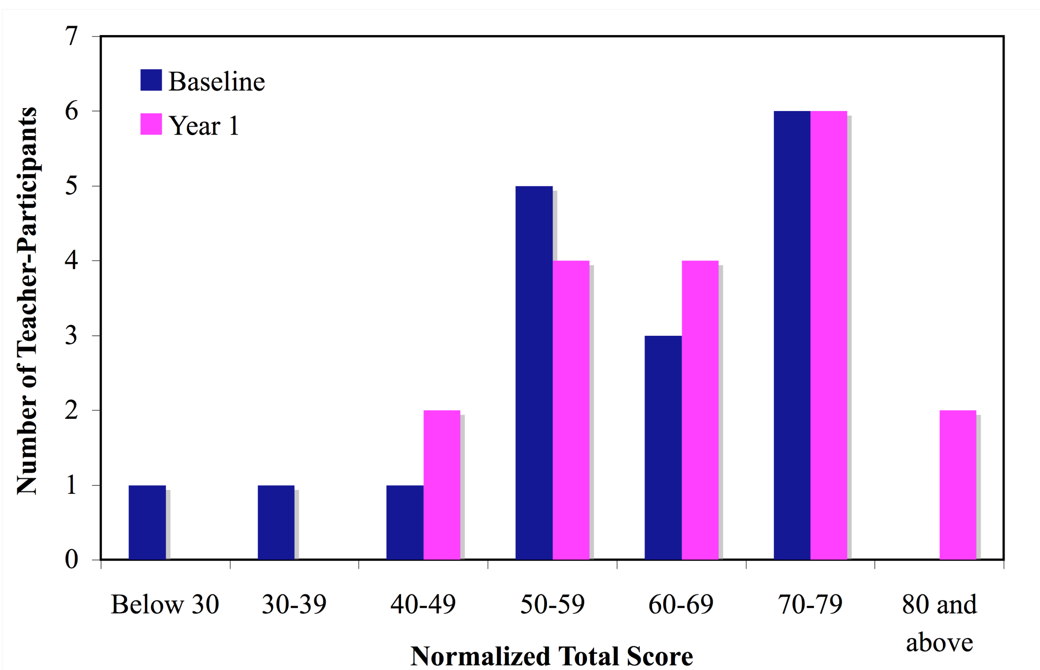

Figure 2. SLPAI score distribution for MSSci Baseline and Year 1 lesson plans.

analysis. These data indicate that the MSSci teacher-participants had made gains in some areas of content and pedagogical knowledge during their first year of Institute instruction, but that certain areas of practice were unaffected thus far.

Table 6
t-test Results for MSSci Teacher Change in SLPAI Categories

SLPAI Category	**Baseline average** (N = 17)	**Year 1 average** (N = 18)	**t-value**
Alignment with Endorsed Practices	68	86	3.38**
Cognitive and Metacognitive Issues	57	66	2.09*
Sociocultural and Affective Issues	68	77	1.66 [a]
Portrayal and Uses of the Practices of Science	49	49	0.009
Total Score	59	65	1.35

[a] Welch correction applied due to non-normal Year 1 distribution.
* $p < 0.05$. ** $p < 0.01$.

MSSci Baseline and Year 1 averages on individual items were also compared to find areas of significant improvement. The items investigated were chosen because they fit at least one of two characteristics: items where the participants' first year MSSci coursework is hypothesized to have an impact, and/or items with a low (< 1.0) Baseline average. Applying repeated measures ANOVA to the raw item score distributions, we found statistically significant score increases on several SLPAI items (Table 7). The alignment of MSSci teachers' lesson plans with state or national standards had improved somewhat, probably due to an emphasis on the need for such alignment in the Science Education course assignment. Similarly, teachers were more likely to attend to changing student attitudes or beliefs about science; however, the Year 1 cohort average score of 0.85 is still below the "Making Progress" level. In contrast, teachers entered the program with a fairly high average score in the category of Classroom Discourse, and still were able to significantly improve this score, perhaps due to the emphasis placed on group inquiry learning in the Science Education course. Finally, teachers had previously researched the literature and interviewed their own students to understand common preconceptions in the topic covered by their lesson plan; this assignment likely accounted for the large and very statistically significant score gain on the Pre-assessment item. Since the cohort average rose from close to zero to well above the “Making Progress” mark, this one item represents the greatest impact of the MSSci course work thus far on the teachers’ practice as measured by the SLPAI. Although other aspects of the study design could account for some of the improvements mentioned here, our ability to directly connect these areas with the teacher-participants’ program experiences allows us to be confident in attributing their participation with the changes described above.

Table 7
MSSci Teacher Change on Key SLPAI Items by Repeated Measures ANOVA (N = 15)

SLPAI item and description	**Baseline mean**	**Year 1 mean**	**ANOVA F value**
Alignment with standards	1.30	1.72	4.73*
Awareness of science education research – reflects knowledge and application of theory	1.40	1.67	2.37
Goal orientation – includes changing student values, attitudes or beliefs	0.00	0.85	61.30***
Pre-assessment – teacher solicits student ideas in order to plan instruction	0.23	1.50	83.02***
Assessment – emphasizes conceptual understanding, includes grading rubric	1.17	1.42	4.20
Equity – attempts to address equity and access for underrepresented populations	0.95	1.10	2.39
Student engagement – motivates students and requires active participation	1.45	1.55	0.31

SLPAI item and description	Baseline mean	Year 1 mean	ANOVA F value
Classroom discourse – fostering a community of learners	1.20	1.70	6.46*
Nature of science – reflects tentative nature of knowledge based on changing evidence, social process involving argumentation	0.60	0.50	0.32
Analytical skills – students learn to support conclusions with appropriate evidence	1.17	1.07	0.14
Analytical skills – the sources and effects of experimental error are discussed (N = 8)	0.28	0.19	0.08

* $p < 0.05$, *** $p < 0.001$

The data in Table 7 also indicate that there were several areas of concern in which teachers began with a low item average score and did not show significant improvement. These include attention to equity, the nature of science, and error analysis. Programmatic efforts to address these areas are ongoing, and preliminary responses by some faculty members will be described below.

Instructor Responses to SLPAI Evaluation Data

As previously mentioned, HSChem teachers submitted baseline lesson plans with low achievement on the nature of science and student inquiry items. MSSci teachers also performed poorly with respect to the nature of science and error analysis items, both in their Baseline and Year 1 lesson plans. Since teacher knowledge, beliefs and practices in these areas are of great importance to the Institute, and relevant to the teachers' Institute content courses, these results were presented to Penn STI science faculty members during team meetings in the spring of 2006. We presented the claim that our teachers "fail to accurately portray science as a process of generating new knowledge, and fail to engage their students in the scientific process," and supported this claim with SLPAI data. The science faculty members were posed the question, “How can the science content courses you teach contribute to helping improve the science-specific aspects of the teaching practices of our participants?” Responses discussed included a more conscious approach to modeling these behaviors as instructors, including the use of more historical information when discussing important scientific concepts, and using inquiry or student-centered teaching methods (rather than teacher-centered, knowledge transmission methods) for content instruction more frequently in Institute courses. The instructors also discussed their own feelings about the importance of understanding and experiencing how scientific knowledge is generated for students of science. Finally, possible reasons for the differences between high-school and middle-school teachers were discussed.

In response to these meetings, several instructors made conscious decisions about instruction. One pair of MSSci co-instructors chose to revise their course extensively, in part to allow time for significant examination of the role of measurement, estimation and error in physical sciences. In the first week of class, students worked in groups to measure a difficult and ill-defined quantity, such as the height of a large statue on campus, and then reported their methods and findings to the class. While planning their procedure, many students asked the instructors to clarify what should be measured, but the instructors left it up to students to define their problem and pointed out that this is always the first step in investigating a scientific question. Before their group presentations, the instructors made explicit that the exercise was a way to experience and understand the role of peer review in the scientific process. Students gave each other feedback about potential problems in their proposed procedures. During the presentations of final results, students often expressed the desire to know the “right answer”, revealing a problematic, naïve realist view of the nature of science typical for many teachers. The instructors responded that "there are no right, definitive answers," put the focus back on analysis and critique of the methods used, and insisted that the students, rather than the instructors, were the arbiters of how "right" an answer was. The messages inherent to this activity were reiterated to students in smaller ways throughout the course. We are interested to see whether this second cohort of students' new appreciation for the roles of uncertainty and peer review in science will translate to their classroom teaching as evident in future lesson plan evaluations.

Another MSSci instructor, whose course had not yet been taught at the time of the faculty meeting, later reported, "your presentation made me think that I wanted to put more (and more explicit) emphasis on the nature of science in my class." She decided to begin her course by emphasizing that students will develop scientific models based on their own observations, rather than formulas from the instructor or textbook. Furthermore, she plans to be explicit with the teachers about the reasons for doing so: to support the development of deeper conceptual understanding as well as appreciation for the process of scientific knowledge generation. The students currently enrolled in this course are the same MSSci teachers used to generate the teacher change SLPAI data presented in this paper; hopefully, their experiences this year will have an impact in the areas of nature of science and error analysis which will be evident in future lesson plans. (We plan to ask teachers to submit Year 2 lesson plans before leaving the program.)

As a group, the HSChem instructors did not respond to the SLPAI data presented to them by considering or making changes to their courses. Since no Year 1 plans were available from the HSChem teachers, we did not have any evidence regarding whether program instruction had an impact on our participants' lesson planning. Given this lack of motivating evidence, the HSChem instructors' reticence can probably be attributed to the fact that the HSChem program is already in its seventh year, and most instructors are resistant to making changes to courses that have been developed and fine-tuned over many years. However, one HSChem instructor described two historical data analysis projects that he has been using for a number of years. The goal of these assignments is to put students in the position of a scientist who has just done an experiment and collected data, and now needs to establish criteria for determining whether the data illustrate a now-known law or mathematical relationship. These verification activities address some of the aspects of the nature of science that students often struggle with: that observations do not usually lead directly to conclusions, and that inferences are accompanied by some level of uncertainty. Feedback from this year's students influenced the instructor plan a class discussion of the project next year, allowing a more explicit treatment of the purposes of the exercise with respect to the nature of science.

Conclusions

We conclude that the SLPAI, which utilizes artifacts of teaching and learning as data sources, is complementary but not redundant to other measures of teaching practice. The SLPAI specifically addresses issues particular to the nature of the science classroom, and is a more easily scalable method than direct observation. An added benefit of lesson plan analysis is that it provides the researcher information about a larger unit of teaching than a one-day observation, offering the researcher a more complete view of a teacher's practices. However, lesson plan review does present some unavoidable sources of imprecision as a measurement technique. A lesson plan, by definition, does not provide information about lesson enactment, unless post-lesson information is also provided. We have also found that evaluators are often more critical of a familiar lesson than one they have not experienced. For this reason, we recommend that evaluators using the SLPAI have classroom experience with the age level and science discipline being evaluated. With these caveats in mind, the SLPAI is a unique and powerful tool for measuring teaching practices over time, especially when used in concert with other measures.

As MSP and other teacher professional development programs expand in result of nationwide efforts to improve teacher quality, especially in STEM fields, evaluation methods that can be used to triangulate other qualitative and quantitative measures will be needed. The SLPAI is an example of such an instrument, which can be used for dual purposes: as a formative tool that informs program development and promotes effective instruction of the teacher-participants, and as a summative measure that allows evaluators to provide a richer, more complete picture of program effectiveness.

References

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academy Press.

Brickhouse, Nancy W. (1990). Teachers' beliefs about the nature of science and their relationship to classroom practice. *Journal of Teacher Education*, *41*(3), 53-62.

Brown A. L. & Campione J. C. (1996). Psychological theory and the design of innovative learning environments: on procedures, principles, and systems. In L. Schauble & R.

Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289-325). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chinn, C. A. & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175-218.

Crowther, D. T., Lederman, N. G., & Lederman, J. S. (2005). Understanding the true meaning of nature of science. *Science and Children, 43*(2), 50-52.

Hacker, R. & Sova, B. (1998). Initial teacher education: a study of the efficacy of computer mediated courseware delivery in a partnership context. *British Journal of Educational Technology, 29*(4), 333-341.

Kahle, J. B. & Scantlebury, K. C. (2006). *Evaluation of University of Pennsylvania Science Teacher Institute – 2005-6.* Oxford, OH: Miami University, Evaluation & Assessment Center for Mathematics and Science Education.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council. (2006). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* Washington, DC: National Academy Press.

Penn STI web site. (n.d.). Retrieved December 14, 2006 from http://www.sas.upenn.edu/PennSTI/MCEP.shtml

Piburn, M., Sawada, D., Falconer, K., Turley, J., Benford, R., & Bloom, I. (2002). *Reformed Teaching Observation Protocol (RTOP): Reference manual.* ACEPT Technical Report No. IN00-3.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, R., Benford, R., & Bloom, I. (2002). Measuring reformed practices in science and mathematics classrooms: the Reformed Teaching Observation Protocol. *School Science and Mathematics, 102*(6), 245-253.

Scantlebury, K., Boone, W., Butler-Kahle, J., & Fraser, B. J. (2001). Design, validation, and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching, 38*(6), 646-662.

Trigwell, K. & Prosser, M. (2004). Development and use of the Approaches to Teaching Inventory. *Educational Psychology Review, 16*(4), 409-424.

Wiggins, G. & McTighe, J. (2001). *Understanding by design*. Upper Saddle River, NJ: Merrill/Prentice Hall.