Development and Use of a Science Lesson Plan Analysis Instrument for Evaluating Teacher Practice and Informing Program Instruction

Christina L. Jacobs & Tracey C. Otieno
University of Pennsylvania Science Teacher Institute

Prepared for

MSP Evaluation Summit II
October 4-5, 2006

ABSTRACT

This paper describes the development and use of an instrument for analysis and evaluation of teacher-submitted multi-day lesson plans. Lesson plan analysis is complementary to teacher survey and classroom observation as a method for measuring changes in the classroom practices of teachers; our purpose is to evaluate the effect of teacher engagement in an intensive professional development program. The *Science Lesson Plan Analysis Instrument* (SLPAI) has been utilized to gain formative baseline information about the teaching practices of incoming program cohorts in order to tailor both pedagogical and content instruction appropriately. The results of a comparative baseline study of incoming high school chemistry teachers and middle grades science teachers indicated several significant differences between the teaching practices of the two groups, which were validated by survey data. Middle school teachers scored significantly higher on average on items dealing with socio-cultural and affective issues and the portrayal and uses of the practices of science. However, both groups had an average score far below the satisfactory level for the individual item dealing with the accurate representation of the nature of science to students. The SLPAI has also been used to track and describe changes in teaching practice and pedagogical knowledge at both the individual and cohort level over time and thereby provide evidence of program effectiveness. After one year of program instruction, statistically significant improvement was seen in the average scores of middle-grades science teachers in the areas of alignment with endorsed practices and cognitive and metacognitive issues. However, no improvements of already low scores were seen on the nature of science and error analysis items. This problematic result was communicated to faculty members, and we report on the responses of several instructors who revised their instructional plans accordingly.

Development and use of a science lesson plan analysis instrument for evaluating teacher practice and informing program instruction

Christina L. Jacobs & Tracey C. Otieno
University of Pennsylvania Science Teacher Institute

Introduction and Purpose

The Penn Science Teacher Institute (STI) 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." In order to evaluate the success of our efforts, valid and reliable measures of teaching practice are needed. Unfortunately, conducting extended observations of all our participants is not feasible for practical reasons; therefore, we have developed and piloted the *Science Lesson Plan Analysis Instrument* (SLPAI) for analysis and evaluation of teacher-submitted multi-day lesson plans. This method evaluates teachers' content knowledge in a classroom context, and gives us a lens with which to assess their pedagogical reasoning (Shulman, 1987). We hypothesize that the results from lesson plan analysis will provide insight into teachers' classroom practices, although the SLPAI is not itself a direct evaluation of teaching practice.

The purpose of this paper is twofold: to present the development and testing of the SLPAI for evaluating teaching practice, and to report on the use of the resulting evaluation data to inform program instruction. We describe the process of SLPAI development and validation, leading to its use in gaining formative baseline information on program cohorts. The instrument was also utilized to evaluate lesson plans after one year of STI participation, in order to track and describe changes at both the individual and cohort level over time and thereby provide evidence of program effectiveness. Data about the teaching practices of program participants were communicated to program instructors during faculty meetings in order to allow instructors to tailor both pedagogical and content instruction appropriately.

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, 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 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 that the item is understood by the teacher in the intended way, such "justification" of students' conclusions could range greatly in quality, reflecting 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, it may not by itself 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; Sawada, Piburn, Judson, Turley, Falconer, Benford & Bloom, 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 direct observation of teachers is resource intense and therefore not highly scalable. In our specific case, our program will have at maximum 120 teacher-participants 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 practice as well; one STI 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 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 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). We utilized 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 STI'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 teacher questionnaire and direct observation instrument utilized in our evaluation, so the SLPAI is a theoretical complement to these methods.

Methodology

Setting and participants

Participants are in-service teachers enrolled in the Penn Science Teacher Institute (STI), funded by the NSF Math-Science Partnership program. The STI is comprised of two Masters degree-granting programs: the Masters of Chemistry Education Program (MCEP) for high school chemistry teachers and the Masters of Integrated Science Education Program (MISEP) for middle-grades science teachers. 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.

Twenty MISEP and eight MCEP teacher-participants were included in this study, although some participants were omitted for certain analyses to allow paired statistical testing. These participants teach in three states and many districts; we evaluated science lesson plans ranging from 5^{th} to 12^{th} grade level. While a large proportion of the teacher-participants work in urban schools, primarily in the School District of Philadelphia, suburban and rural schools are also represented in the sample. The teachers have a wide range of experience levels (two to twenty-five years teaching). They are also diverse with respect to their prior science backgrounds, having taken from one to fourteen post-secondary science courses prior to enrolling in MISEP or MCEP.

Instrument development and testing

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

The SLPAI consists of four major categories: Alignment with Endorsed Practices (AEP), Lesson Design and Implementation – Cognitive and Metacognitive Issues (CMI), Lesson Design and Implementation – Sociocultural and Affective Issues (SCAI), and Portrayal and Uses of the Practices of Science (PUPS). The full list of item titles by category is given in Table 1. For each item, teachers could be rated as Exemplary (2 points), Making Progress (1 point), or Needs Improvement (0 points). These scores were then multiplied by the weight of the item (ranging from 1-3), 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 affecting the overall score. Weights were determined according to the goals of the STI, and are meant to provide flexibility in adapting the SLPAI to other contexts.[1]

The reliability of the SLPAI was examined using double-scoring approximately 20% of the lesson plans by the co-developers of the instrument. The average inter-rater agreement in this test was 96%. Training of an outside researcher and further reliability testing is currently underway.

Baseline diagnostic pilot study

Prior to beginning their STI programs, teacher-participants were asked to submit a sample unit lesson plan of approximately 5 days in length. 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 teacher-participants, 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 MISEP plans and 8 of 14 MCEP 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. These teachers were omitted from the analysis.

[1] A full version of the SLPAI is available from the first author to program evaluators and teacher educators upon request.

TABLE 1
SLPAI items by category with scoring weights

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

Teacher change pilot study

MISEP teacher-participants 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 "Year 1" plans, which covered 2-3 days of instruction, and they were scored using the SLPAI. Total and sub-score distributions were compared to the Baseline scores for significant change using unpaired t-tests (17 Baseline, 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 (15 of 21 teachers).

Because the course assignment that was provided the Year 1 data did not entail submission of graded student work, or require that the lesson had been implemented in a classroom, analysis of these plans could not provide any information about lesson enactment. Furthermore, these lesson plans were submitted for a grade, unlike the Baseline lesson plans. 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 RTOP (Sawada et al., 2002), and the *Standards-Based Teaching Practices Questionnaire* (SBTPQ, Scantlebury et al., 2001).

Results and Discussion

Validation of the SLPAI

Two cohorts of teacher-participants 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 MCEP and MISEP teachers as a means for instrument validation.

While the SBTPQ and SLPAI emphasize different aspects of the teaching endeavor, we found that conclusions drawn from the STBPQ were supported by the SLPAI results. For example, using the SBTPQ, external evaluators found that MISEP teachers reported significantly more frequent use of standards-based teaching practices than the MCEP teachers (Table 2, taken from Kahle & Scantlebury, 2006), both in terms of what they do in class and what their students do. MISEP teachers were significantly more likely than MCEP teachers to arrange seating to facilitate student discussion, use open-ended questions, require students to supply evidence to support their claims, encourage students to consider alternative explanations, and use non-traditional or authentic assessments. MISEP teachers also reported that their students were more likely than MCEP teachers' students to design activities to test their own ideas and talk with one another to promote learning.

The conclusions drawn from the SQTPQ measure were supported by the baseline SLPAI data. Four SLPAI items were identified to address the same teaching practices listed in Table 2, and the Baseline results for MISEP and MCEP teachers in these four items and their respective subscales were analyzed using unpaired t-tests for significant differences between cohorts. We found that MISEP teachers scored significantly higher than MCEP teachers on all items and subscales tested, as shown in Table 3. The largest average score difference between MISEP and MCEP teachers was in the promotion of active student engagement ($p < 0.01$).

TABLE 2
SBTPQ items and subscales with significant baseline differences for MISEP and MCEP teachers (1-5 point scale)

SBTPQ Item or Subscale	**MISEP (N = 23)**		**MCEP (N = 18)**		**t-value**
	Mean	**SD**	**Mean**	**SD**	
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*
"In my classroom" subscale average	3.52	0.48	3.08	0.29	11.70**
"My students" subscale average	3.32	0.38	3.05	0.40	4.85*

* $p < 0.05$, ** $p < 0.01$

TABLE 3
Baseline results for MISEP and MCEP teachers on SLPAI items (0-2 point scale) related to the SBTPQ items in Table 2

SLPAI Item	**MISEP (N = 17)**		**MCEP (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 sense-making discussion among students, open-ended discussion questions are provided	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*
Sociocultural and Affective Issues subscale (normalized)	53	21	49	19	2.13*
Portrayal and Use of the Practices of Science subscale (normalized)	36	19	29	19	2.47*
SLPAI total score (normalized)	47	14	46	14	2.21*

* $p < 0.05$, ** $p < 0.01$

Validation of the SLPAI using direct classroom observations and RTOP data is ongoing. We are unable to report any findings in this area at this point, in part because of the small number of teachers observed using RTOP thus far.

In addition to the areas of overlap between the SLPAI and other evaluation instruments, there were also aspects of teaching practice that are best addressed by the SLPAI. The SLPAI includes an item that examines in some depth how a teacher represents the nature of science to students: as a social endeavor that is evidence based, iterative, dependent on indirect reasoning, and subject to debate and change, rather than as a body of facts to be memorized and verified through unproblematic "cookbook" laboratory activities. While the SBTPQ does include items concerning with relevance to the nature of science, such as "I require that my students supply evidence to support their claims," "My students argue or debate with one another about the interpretation of data," and "I discuss experiments from the history of science," the teachers' responses to these questions did not allow the external evaluation team to draw any significant conclusions about how our participants represent the nature of science in their classrooms. In contrast, the SLPAI's more detailed look at nature of science issues provided clear evidence that teachers in both programs do not give their students an accurate picture of how scientific knowledge is generated and tested. One possible reason for the lack of information generated through survey responses lies in the link between a teacher's epistemology and his or her response to SBTPQ items like those listed above. Since survey questions are interpreted by the teacher (rather than lesson plans interpreted by the researcher), a teacher's non-normative view of the nature of science may lead to an unfounded positive survey response due to misunderstanding of the item's intent. For example, discussion of historical experiments may provide students with an accurate idea about how knowledge is generated, but only if those epistemological ideas are supported in the discussion. Teachers may also discuss historical experiments in class for other purposes, or they may impart a normative view of the nature of science without discussing historical experiments – the observed variable (survey response) and variable of interest (teaching about the nature of science) do not necessarily co-vary in this example.

Turning to RTOP, the only mention of nature of science issues is a criterion that asks the observer to rate the extent to which "intellectual rigor, constructive criticism, and the challenging of ideas were valued" within the classroom. We believe this to be an oblique reference to the nature of science, and unlikely to give much information about changes in teacher practice in this area. Since the RTOP was developed for use in both math and science classrooms, this item lacks the specificity and clarity of the "Nature of Science" item from the SLPAI.

An important aspect of teaching practice, and one not addressed by the SBTPQ, is that of teacher knowledge. As discussed above, teachers' understandings of a content discipline, and how to best teach it, directly and indirectly affect the quality of the instruction they provide their students. With this in mind, we included two items dealing with the teacher's understanding of and presentation of content: the "Content accuracy" and "Content presentation" scales. While the RTOP does include a subscale concerned with "Propositional knowledge", observation of a single class may not always provide insight into the multitude of content-based decisions a teacher makes. One example is the determination of what counts as a correct or incorrect solution to a problem; such decisions are most easily identified and evaluated in the context of graded student work,

which can serve as a useful indication of the limitations in a teacher's knowledge. Furthermore, the longer timeframe covered by a unit lesson plan allows evaluation of the sequencing of topics and instructional activities, which is also a key component of effective content presentation and a marker for sophisticated pedagogical content knowledge.

Baseline diagnostic pilot study

Baseline lesson plans were analyzed using the SLPAI in order to provide information on the strengths and weaknesses of incoming teachers' knowledge and practices. Table 4 presents the results for all items of interest; these are defined as items with either low or high cohort averages (less than 1.0 or greater than 1.5, respectively) and/or items with a significant difference in MISEP and MCEP cohort averages. Only items with scoring weight greater than one are included, since these are the areas of most relevance to the STI.

From these data, we see that both cohorts were very strong in the areas of content accuracy and content presentation. 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 to these commonalities between the cohorts, MISEP teachers also performed well on the item dealing with student engagement, but poorly on the pre-assessment item. MCEP teachers did not attend to the issues of classroom discourse, variety, and student inquiry at a high level in their lesson plans.

In addition to the four items discussed previously (Table 3) in the context of instrument validity (student engagement, classroom discourse, student inquiry and analytical skills), there were statistically significant differences in cohort performance in several other areas. MISEP teachers scored higher on average than MCEP teachers on items concerning meaningful application of science content, student reflection, variety, and nature of science. (Note, however, that the average MISEP score in the nature of science category was still well below 1.0.) On the other hand, MCEP teachers outscored MISEP teachers in the area of goal orientation.

These data suggest that teachers enter our program with a foundation of practical knowledge and experience that can be built from. 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 presented in their Baseline lesson plans. Although on the surface this result might be read to suggest that the STI's focus on improving content knowledge of secondary science teachers is unfounded, we believe that the intensive science coursework provided in the STI will enable teachers to expand their comfort level with science, improve the accuracy of their teaching, 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 many areas, our teachers' lesson plans do not include evidence of social constructivist teaching practices and beliefs. This result points to the need for the STI programs to address educational theory and the link between content knowledge as learned in the program and as utilized in the secondary classroom. Finally, we identified several areas in which MISEP and MCEP teachers performed at different levels; many of

these were consistent with previous findings from self-report survey data described above. Importantly, the areas of goal orientation, application, variety, and the nature of science were only diagnosed as baseline differences by the SLPAI (and not survey data); this indicates the instrument's usefulness in uncovering useful information about teacher knowledge and practices.

TABLE 4
Item analysis of baseline SLPAI results by program (0-2 point scale)

SLPAI Item	**MISEP (N = 17)**		**MCEP (N = 8)**		**t-value**
	Mean	**SD**	**Mean**	**SD**	
Goal orientation – learning goals are explicit, comprehensive and fundamental, and are supported by learning activities	1.12	0.34	1.46	0.40	2.53*
Content accuracy	1.66	0.58	1.69	0.53	0.75
Content presentation – level of detail and abstraction are challenging but accessible, sequence is appropriate, appropriate examples are included	1.54	0.49	1.53	0.67	0.70
Pre-assessment – teacher solicits student ideas in order to plan instruction	0.32	0.56	N/A‡	N/A	N/A
Meaningful application – content is given a personal or real-world significance to students	1.31	0.58	0.56	0.42	3.52**
Student reflection – students reflect on and summarize their understanding	1.19	0.65	0.71	0.55	2.17*
Variety – Teacher innovation or creativity keeps teacher and students engaged	1.47	0.57	0.97	0.66	2.29*
Nature of science – tentative nature of knowledge based on changing evidence, social process involving argumentation	0.68	0.52	0.00	0.00	3.66**

* $p < 0.05$, ** $p < 0.01$, ‡The "Pre-assessment" item was added to the SLPAI during a later round of revisions, and after MCEP Baseline lessons had been evaluated.

MISEP teacher change pilot study

At a coarse level of analysis, the total score distributions of the MISEP Baseline and Year 1 lesson plans are represented below in Figure 1. 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 5).

No change was seen in the PUPS category although it had the lowest Baseline category average. Note that this comparison is between non-paired scores, 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 analysis.

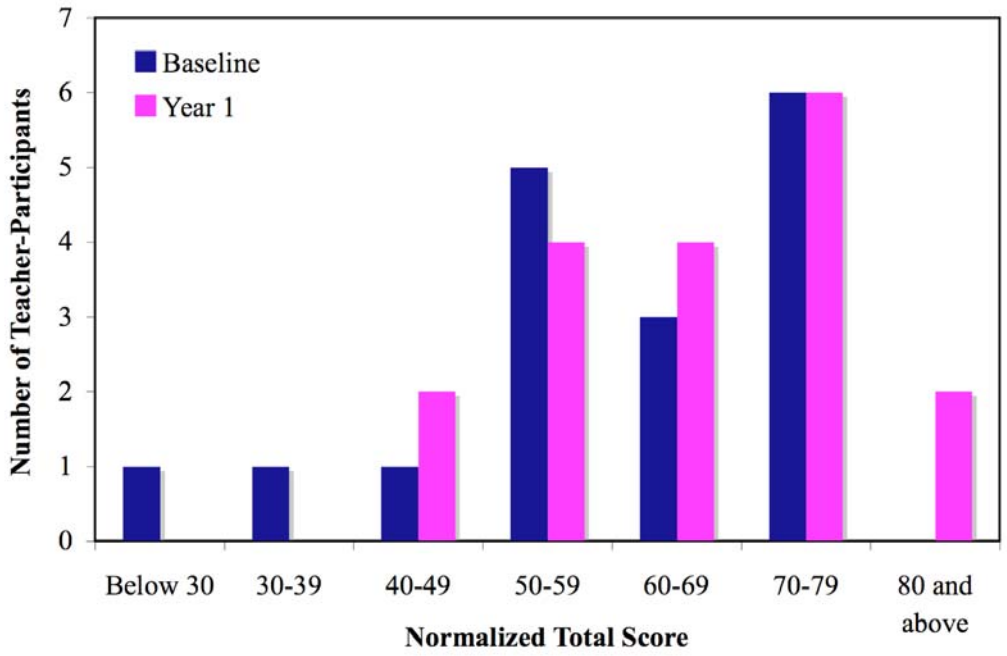

FIGURE 1. *SLPAI Score distribution for Baseline and Year 1 lesson plans*

TABLE 5
T-test results for MISEP 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‡
Portrayal and Uses of the Practices of Science	49	49	0.009
Total Score	**59**	**65**	**1.35**

* $p < 0.05$, ** $p < 0.01$, ‡Welch correction applied due to non-normal Year 1 distribution

MISEP 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: 1) items where the participants' first year MISEP coursework is hypothesized to have an impact, and 2) 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 6). The alignment of MISEP 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 area 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 MISEP 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 6
MISEP teacher change on key SLPAI items, using repeated measures ANOVA (N = 15)

SLPAI Item	**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
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 6 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, MCEP teachers submitted baseline lesson plans with low achievement on the nature of science and student inquiry items. MISEP 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 STI, and relevant to the teachers' STI content courses, these results were presented to 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 courses you teach (as opposed to the science education courses) 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 STI 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 hypothesized.

In response to these meetings, several instructors made conscious decisions about instruction. One pair of MISEP 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 MISEP 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 MISEP 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 MCEP 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 for the MCEP teachers, we do not have any evidence at this point regarding whether program instruction has an impact on our participants' lesson planning. Given this lack of motivating evidence, the MCEP instructors' reticence can probably be attributed to the fact that the MCEP 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 MCEP 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 using the SLPAI to measure teaching practice is complementary to the use of teacher and student surveys and direct observation using RTOP but not redundant, since each method uses a different lens with which to view teaching practice. The SLPAI addresses issues particular to the nature of the science classroom, and is a more easily scalable method than direct observation using RTOP. An added benefit in lesson plan analysis is that it allows consideration of a larger unit of teaching than a one-day observation does, thereby offering the researcher a more complete view of a given teacher's practice. The SLPAI represents a compromise between the rich but expensive data than can be generated through direct and prolonged observation of teachers and classrooms, and the plentiful but indirect data that result from large-scale survey deployment.

However, while developing and utilizing the SLPAI during the past year, we realized that lesson plan review does present sources of error. Limitations of our study design resulting in incomplete data sets and non-ideal comparisons between Baseline and Year 1 data have already been described. In addition, since lesson plans are indirect sources of evidence for classroom practice, it is difficult for the evaluator to avoid interpretation of the lesson plan through the lens of their own experience and knowledge. We noted that we were more comfortable evaluating a teacher's treatment of topics that we were more familiar with ourselves, so that our ratings disagreed more when one of us had significantly more knowledge in the subject than the other. This source of rater disagreement has also been noted when using RTOP, although to a lesser extent.

The design of the rubric was intended to allow program instructors to adapt the applicable portions of the rubric for grading the participants on course lesson-planning assignments. This approach will promote a coherent, program-wide approach to strengthening the pedagogical knowledge of teacher participants. Two different MISEP courses utilized the SLPAI as a grading rubric for lesson planning assignments this summer; whether or not this experience has a lasting impact on the participants' lesson planning will be investigated in future rounds of evaluation.

In summary, SLPAI data from two pilot studies were used to diagnose areas of strength and weakness of incoming cohorts, as well as significant differences between the two populations of teachers. One cohort was evaluated a second time, in order to test the usefulness of the SLPAI in measuring teacher change, and to generate a mid-program data set in order to monitor program effectiveness. This study found that teachers' practices as reflected in their lesson plans improved in a few areas, and these improvements could be attributed to aspects of the science education course they were enrolled in at the time. However, there were several areas in which teachers were deficient and remained so after a year of STI participation, including how the nature of science was represented in their lessons. These findings were communicated to science faculty members, some of who responded to this impetus by planning or re-examining how they would engage students in scientific inquiry and represent the nature of science in their own STI classrooms. It is important to note that the most important result from this study with regard to program instruction involved nature of science issues; utilizing only the SBTPQ and RTOP instruments, it is unlikely this deficiency in our teachers would have been diagnosed and addressed. This demonstrates the usefulness of lesson plan analysis as a program evaluation tool, and of the SLPAI as an evaluation instrument.

We are continuing to collect SLPAI, SBTPQ and RTOP data on the cohorts described in this study, as well as incoming cohorts of STI participants. A growing data set will allow additional instrument validation, including the comparison and alignment of direct observation data with lesson plan analysis. We plan to continue using such data as a means of feedback to STI instructors. In particular, as an institute that serves a large number of urban schools, we would like to review data regarding equity in the secondary science classroom with our science education faculty, with the aim of improving our teachers' abilities to support science learning for all students. We intend to follow up on the changes STI faculty have made in their courses, and hope to see score improvements on the relevant items and categories of the SLPAI as a result of these changes. We are optimistic that the use of a targeted instrument like the SLPAI, in conjunction with ongoing program review and faculty development meetings, will assist our programs to

reach their goals of “increasing the content knowledge of science teachers, and changing the teaching and learning methodologies used in science classrooms to research-based promising pedagogical practices.”

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.

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.

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review, 57*(1), 1-22.

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.