PAPER SESSION

MOSP: From MSP to Teacher to Student

CONTENTS

Session Presentation .................................. 3
Test Construction ..................................... 6
What Is the State of Student and Teacher Knowledge? .... 6
Patterns in Test Data .................................. 7
Research Uses of the ASCII Instruments ................. 10
Results of Teacher Professional Development ...... 12
Next Steps: Attribution of Gains to Institute Attributes .... 13
Conclusions, Findings, Key Features .... 14

Participant Q & A ..................................... 16
About This Summary

This documentation of the 2011 Math and Science Partnership Learning Network Conference offers a brief summary of the presentation that took place during one conference breakout session and focuses on questions, answers and discussions during the session.

The abstract for this presentation is posted on MSPnet at the following URL:


The color PowerPoint presentation is posted on MSPnet at the following URL:


More from LNC 2011 on MSPnet

For a summary of LNC plenary session proceedings, as well as video of those sessions and all conference abstracts and PowerPoint presentations, visit:

www.MSPnet.org

This documentation is part of the MSPnet project, funded by the National Science Foundation and designed and facilitated by TERC Inc., a not-for-profit education research and development organization based in Cambridge, Massachusetts.

Participant comments have been paraphrased; they are not exact quotes. The contents of this document do not necessarily reflect the views of TERC, the National Science Foundation, or the organizations of any participants.

www.mspnet.org

An electronic learning community for all MSP projects.

Principal Investigators
Joni_Falk@terc.edu
Brian_Drayton@terc.edu

Conference Documentation
Catherine McEver
Phillip Sadler opens the session by noting that as a RETA project, MOSART looks at what is happening in the entire MSP system rather than what is happening within an individual project. The big question is: How do you rigorously measure conceptual understanding of teacher and students? A goal for most MSPs is to increase teacher subject matter knowledge and/or pedagogical content knowledge, and in turn to see an increase in conceptual understanding on the part of students, Sadler notes. You could also measure their persistence in the field, he observes, such as whether they want to enter a STEM profession or want to take more STEM courses. MOSART has focused on conceptual understanding of teachers and students, and project members have written extensively on this subject and what it takes to measure conceptual understanding in a valid way.

Sadler shows a video demonstrating the starting point in this process, which involves sitting down with students and asking them a question that relates to a concept in science. These concepts are often part of the NRC standards. In this example, students are building scale models of the solar system and answering questions about what they are doing.

How do you rigorously measure conceptual understanding of teachers and students?


Sadler points out that the interviewer doesn’t give away any right answers, accepts the student’s comments and asks for definitions of terms used, such as “scale model.” Kids as young as ten understood what a scale model was and could identify scale and non-scale models, he reports.

The challenge is how to move on from these interviews that MOSART and others have done. There are over 1,000 papers on children’s ideas regarding the sciences. How, asks Sadler, do you go from these qualitative studies to assessments that actually measure something important?

Both students and teachers have (or had) preconceptions

- Exist prior to formal instruction
- At odds with accepted scientific thought, “misconceptions”
- Commonly held, not idiosyncratic
- Embedded in larger knowledge structures, not just a simple “error” (that is easy to correct)
- Resistant to change, overestimation of ∆
- Originally catalogued as “math bugs”
- Preconceptions forgotten as knowledge is restructured

Part of this involves recognizing the role of preconceptions, and that these preconception exist prior to formal instruction. In middle school physical science, for example, students arrive with many preconceptions that they’ve gotten from other places and are at odds with scientific thought. They are commonly held, which means there are just a few for every topic. These are embedded in larger knowledge structures, they are not simply an error or problem that can be corrected easily, Sadler explains. They are manifestations of a knowledge structure that remains hidden, like the tip of the iceberg. They are resistant to change and teachers generally vastly overestimate any change in the student, Sadler reports. Students may be able to answer a question correctly on a teacher-made test, but their conceptions have not shifted.

One of the more interesting phenomena is that we forget the misconceptions we all had, Sadler notes. We remember the science and assume we never had these misconceptions or had very few of them. In reality, he states, we all went through this process.

MOSART approached this issue by taking all of

Steps in instrument development based on student ideas

- Employ NRC standards, the root of state standards
- Construct assessment instruments based on misconceptions
- Research literature
- Validation with both students and teachers
- Pilot and field tests
- Final instruments
- Measure both SMK and PCK
the National Research Council Science Standards and converting them to multiple choice items that embed misconceptions. The test currently includes about 2,000 items, which have been validated with large groups of students.

For comparison, Sadler offers a NAEP item that addresses the cause of day and night. In this nationally representative third grade group, 81% got the answer right. A teacher or researcher might look at this and think that 81% is pretty good and the students have mastered this by third grade.

What is the cause of day and night?
NAEP 3rd grade
a. Earth turns. (81%)
b. Sun turns. (8%)
c. Moon turns. (4%)
d. Sun gets dark at night. (6%)
e. I don’t know. (1%)

“From our point of view there is something else going on,” Sadler observes. MOSART tested a similar item in eighth to twelfth grade earth sciences and astronomy classrooms.

What is the cause of day and night?
SED 8-12th grade
a. The earth spins on its axis.
b. Clouds block the sun’s light.
c. The earth moves into and out of the sun’s shadow.
d. The sun goes around the earth.
e. The earth moves around the sun.
f. I don’t know.

There is an equivalent correct answer (a), but the distractors (the wrong answers) are different and are misconceptions documented in the literature. Just 68% of the students got the correct answer. There are two ways to interpret this, Sadler notes. Either kids get stupider over time, or there is some intrinsic difference in this kind of question.

The NAEP and many other tests, including the state assessments, don’t develop items based on the literature, Sadler points out. If your test had items like the NAEP example you might conclude that kids had learned a lot, while if you administered the second instrument you might not see the same size gains.

It breaks down as shown at right for this item, with most kids in grades 8-12 getting it right and about a quarter getting the same wrong answer. To explain why you don’t see items like this, Sadler points to the graph at bottom right, the type of graph psychometricians generally use to determine whether standardized test items work well or not. It turns out that the kids who have a moderate level of knowledge actually do worse on this item, Sadler observes, and when psychometricians at other places see this they’ll say, “Terrible item, we can’t use it.
Our Criteria for Conceptual Understanding

Students and teachers must:
- Prefer accepted scientific explanations over widely-held misconceptions
- Item must contain both
- Apply their knowledge to make accurate predictions
- Can concept be used?

Test Construction

Sadler then shares the steps involved in the long process of test construction. There are two sets of tests, one publicly available and a second set that is secure and only for research purposes.

Test Instruments Available at the MOSART Web Site

<table>
<thead>
<tr>
<th>Subject</th>
<th>K-4</th>
<th>5-8</th>
<th>9-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astronomy</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Earth &amp; Space Science</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Physical Science</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Chemistry</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Life Science</td>
<td>S '11</td>
<td>S '11</td>
<td></td>
</tr>
</tbody>
</table>

Sadler proceeds to share what has been learned from these tests, working with upwards of 80,000 kids and thousands of teachers, regarding the state of teacher and student knowledge. This is measured for each grade band for each subject. The first graph (top right, page 7) shows the performance of fifth grade students on K-4 standards, so presumably they would have been exposed to these standards in elementary school. Their performance averages around 50%. The graph also indicates that teachers did fairly well on the same test instrument, so teachers did fairly well while students did poorly. Finally, the graph shows teachers’ estimations of students’ test performance based on how well students do in their own classrooms. There is a universal overestimation of student performance, which we also see in subsequent graphs.

On the next graph we can see that middle school teachers are a little weaker on middle school astronomy standards. In the third graph, teachers do very well at the high school level, but are still overestimating their students’ scores, and the students don’t do well at all.

Results are similar in the physical sciences (though with more accuracy in prediction in the
middle school physical sciences), in high school chemistry, and high school physics, Sadler reports. On the middle school life science field test the teachers themselves perform very well.

Nancy Cook Smith steps in to talk about item-level patterns in the test data. She begins by noting that MOSART does a lot of analysis, but the most familiar is a classical approach to test item analysis.

On the X axis in the graph shown in the sidebar on page 8, the difficulty of the item is expressed as the fraction of students who answered the item, which ranges from no students (for the most difficult item), up to 75% of students answering an item correctly. The Y axis is the discrimination, which is the correlation between the correct response, right or wrong, with the total test score, such that when the discrimination is perfect it divides the upper level from the lower level perfectly, Smith explains. If you look at the .5 on the X axis, you can see that the discrimination is going up and the parabolas show that the three groups—elementary, middle and high school—follow a similar pattern in that the better discriminators

See the slide presentation available on MSPnet for larger color versions of these and other graphs used to illustrate this session: http://hub.mspnet.org/media/data/46_slides.pdf?media_00000007133.pdf
are going to be the ones that are moderate in difficulty.

The interesting thing, Smith points out, is that some of the hard questions are hard because they have the misconceptions embedded in them. She adds that they see a different pattern than most classically oriented tests that try to have a bell curve. “Good discrimination is not necessarily the sine qua non of our tests,” Smith states. “We like to see the range of difficulty. We don’t want to see poor discriminators, necessarily, but we don’t really push the discrimination of the items that much.”

Another characteristic unique to this research is a statistic they’ve named “misconception strength.” That is, of the students who answer incorrectly, how many of them pick the one distractor that reflects the most commonly held misconception? In the three graphs grouped at lower left you see there is not much relationship between the strength of the misconception and the difficulty of the item. “We think this is another indicator of item quality that we use to choose the items in our test,” Smith explains. “The benchmark for misconception strength is, if more than half of the students who are incorrect choose one distractor (and remember there are four incorrect choices), it is deemed a good item for us.”

Phillip Sadler steps in to explain the graph below, showing results of an experiment that involved removing the most popular misconceptions from items and putting in some other answer. “This is a demonstration that if you eliminate the misconceptions, the kids do much
better on the test,” he explains. “They’re not just guessing at the answers, they’re actually choosing this particular wrong answer for a reason.”

The next graph (below left) depicts what happens when the test is given to teachers. Each dot is a test item and the items concern concepts at the grade level that they teach. This show items where all of the teachers got the items correct, but the student performance was only fifty percent at the end of the year. What this means, Sadler explains, is that if you’re trying to get teachers’ subject matter knowledge to be perfect, you still may get only a fifty percent rate among the students. “Teacher subject matter knowledge is not the only thing that’s important here.”

Looking across grade bands in astronomy, student performance is 45% of where it should be, and teachers uniformly over-estimate student knowledge, Sadler observes.

In the graph at the bottom right, each dot is an item. On items that are easier for students, teachers under-predict a bit, but as the items get harder, teachers dramatically over-predict how they think their students will perform.

The next graph is one that Sadler identifies as a favorite. Each dot is a teacher and the
The horizontal axis indicates subject matter knowledge. The vertical axis indicates how well teachers were able to predict the most common wrong answer. Teachers do very well on subject matter knowledge, but can only pick the most common misconceptions about half the time. “So their pedagogical content knowledge as related to their knowledge of misconceptions is quite low,” Sadler observes. “Now what does that have to do with anything? Certainly subject matter knowledge is important, but is there a role for this pedagogical content knowledge regarding misconceptions?”

Research Uses of the ASCII Instruments

Turning to results, the graph at left below indicates the results of a year of instruction in astronomy for middle school students, showing grade bands K-4 and 5-8. The gain at the K-8 level is about three percent which, Sadler notes, is about twenty percent of the standard deviation. There were small gains at the middle school level on middle school standards and insignificant gains on standards that are lower.

At the high school level, shown in the graph at lower right, we see fairly large gains in astronomical knowledge by astronomy and earth science students at the high school level, about two-thirds of a standard deviation; no gains at the middle school level of standards; and small gains on the K-4 standards, which are observational.
Upon analysis, Sadler notes, there are two types of questions on these tests (see graph below right). There are questions in response to which there are strong misconceptions, where over 50% of kids choose one particular wrong answer. There are items without strong misconceptions, which may be more factual or where there’s no impediment from a misconception in learning something. The project then looked at the teacher knowledge: did the teacher get the item wrong; did the teacher get the item right; or did the teacher get it right and know what the misconception was?

The x-axis on this graph is effect size: 25% is a small gain, 50% is a medium gain and 75% is a large gain. This collapses the pre- and post-tests into gains, Sadler explains. The model includes lots of controls, he adds, including student parental education, school SES, gender, “all the usual subjects.”

When the teacher doesn’t know the science, when they get the item wrong, there is still a small gain in both groups, Sadler observes. “The students learn even if the teacher doesn’t get it.” However, he adds, if the teacher knew the science for factual items like the number of planets in the solar system, the gain was three times bigger.

This is also controlling by classroom, Sadler explains, with teachers serving as their own control. Pretty much all of the teachers got some items wrong, he elaborates, so what this graph shows, aggregated across this huge data set, is what happens in a classroom where the teacher knows the science and where the teacher doesn’t know the science.

Another point illustrated by this graph is that if there is a strong misconception, the fact that the teacher has the science knowledge makes very little difference. The difference between the teacher who knows the science and the teacher who doesn’t know the science is very tiny. However, if the teacher happens to know the misconception, the gains are larger.

“So you get moderate levels of gain independent of looking at what the teacher does in the classroom,” Sadler observes. “This is just what the teacher knows. Independent of size of classroom, use of inquiry, or anything like that, for the items where the teacher knew the misconception the gain was fifty to sixty percent larger. The emphasis that we have in MSPs on increasing subject matter knowledge is probably a good thing, but it may only show up on concepts for which there are no misconceptions. It looks like it is very important for teachers to learn what the misconceptions are.”
Regarding teacher development, the project has also looked at several institutes. The two graphs at left below depict pre- and post-test results for a one-week and a two-week astronomy institute. Teacher above the line gained and those below the line diminished in their understanding.

Results from the one-week institute show few or no gains, and the worry was that there was a ceiling effect with high-performing teachers, but other institutes like the two-week institute depicted here, which also had high-performing teachers, did have gains. This provides a way of pre-and post-testing teachers and comparing results.

You can look across institutes, as in the graph below, which shows two institutes with moderate gains in subject matter knowledge and two where it was not significant. “Just because faulty are involved or teacher educators are involved or teachers get together for a week or two weeks and work on something does not mean that there is subject matter knowledge gain,” Sadler points out.

---

1-Week Astronomy Institute
- Learn to use professional instrumentation
- Earth-Sun connection only
- Disciplinary domain focus
- Only relevant items
- Speakers
- Activities
- Observing

2-Week Astronomy Institute
- Basics
- To boost astronomy background
- General astronomy test
- Speakers
- Activities
- Observing

High initial knowledge
No gains at highest level of teacher knowledge
Many teachers with no or negative growth

Moderate initial knowledge
Gains at all levels of teacher knowledge
Few teachers with no or negative growth
Sadler describes this next step, which looks at teacher views of institute attributes, as more of a pilot. The interest here is less in whether institute A does better than institute B on some measure, but rather in the factors that result in larger gains.

The graph at left below shows teacher views of what was effective in the institutes. For example, they thought the teaching techniques were fairly effective. However, the lowest-rated factor here is “I found my ideas about how to teach life science changing.”

Even more interesting, Sadler notes, are the differences by institute. If you aggregate across institutes, what is the variation in institutes?

In the graph at right below, the x-axis includes elements that professional development staff say make them different from everybody else, such as workshops by scientists, field trips, and cutting edge science. The y-axis is the teachers’ assessment of how much of each activity was in the institutes they attended. If you look at an activity like “lectures and workshops by scientists,” there is not much variation. The standard deviation is quite small around that data point; almost all institutes have lectures by scientists. There is larger variation with activities like collaboration with colleagues at the institute or taking field trips or conducting research in science. “So these become candidates for regression models, looking at institutes across the United States.” Sadler explains. “Do field trips help? That is something that we will be able to answer, which an individual institute might not be able to answer.”

Participant Comment

- You have combined “lectures and workshops by scientists” as one element on this graph. I would think that’s something you’d want to tease apart in the professional development, so I’m a little concerned. • Participant

- This was from our pilot with about eight institutes in middle school life science. We have a study coming up with 60 professional development programs. • Phillip Sadler
This graph shows gains in units of effect size on subject matter knowledge in this pilot research involving a small number of institutes and 113 teachers. In middle school life science, several institutes had significant gains in their life science knowledge, but there were many institutes where no gain was measured in teacher knowledge. Teachers still had some holes in their knowledge of middle school life science.

More surprising is the next graph, which indicates measurement of the teachers’ pedagogical content knowledge. There were some institutes where there was a gain, but there were several institutes where teachers were worse at predicting what the most common student answer was. “So by no means are all teacher institutes improving teacher knowledge in all ways,” Sadler notes.

Sadler closes the presentation by pointing to conclusions, findings and insights to date.

Assessment and PD
For each standard at each level
- Students have not achieved mastery
- Teachers generally overestimate student knowledge.
- Teachers know far more than their students
- Teacher knowledge is a not a guarantee of student knowledge
- Subject do much better on items if misconceptions are not a choice
- SMK and PCK not highly related to each other
  - Both predict performance
- PD can have varying degrees of effectiveness
  - Advanced activities have little effect on basic conceptual understanding
  - Experience has little impact on understanding student difficulties

Key Findings
- Institutes vary in the pre-post gains made by teachers in both SMK and PCK.
- Length of time spent on various participant activities, rather than more formal (passive) learning, improves both SMK and PCK.
- Students taught by teachers who have SMK are more likely to change their conceptions than students of teachers who do not know the correct response.
- Students of teachers who accurately predict the most common incorrect response to an item are more likely to reject that conception and embrace the scientifically correct response.
Key insights that have value for the Learning Network

- With our new online testing system, we expect to provide early useful pre-test results to participating MSPs. The first report will be a “diagnosis” of participants’ areas of strength and weakness.
- The use of MOSART assessments will allow us to compare the gains made by one institute’s participants to other institutes targeting the same content and thus facilitate the sharing of practices.
- Using robust, technically validated assessment data minimizes the problems of “apples and oranges” in comparing MSPs. In addition, entirely idiosyncratic measures are often qualitative, and may be biased, as well as difficult to evaluate. With a common evaluation suite, NSF can gain better understandings of what MSPs are accomplishing.

Publications


Acknowledgments

- Co-investigators: Matthew Schneps, Roy Gould, Gerhard Sonnet
- Project Manager: Hal Coyle
- Survey Staff: Jaimie Miller, Nancy Cook Smith, Cynthia Crockett, Marc Schwartz (McGill), Annette Trenga, Bruce Ward, Bruce Gregory
- Graduate Students: Zahra Hazari, John Loehr
- Advice: Elizabeth VanderPutten, Janice Earle, Joyce Evans, Barry Sloane, Larry Suter of the National Science Foundation
- Financial support Smithsonian Institution, NSF, NIH, NASA Annenberg Foundation.
- Center for Astrophysics Irwin Shapiro, Susan Roudebush, Judith Peritz, Charles Alcock.

MOSART Web Site

www.cfa.harvard.edu/smgphp/mosart

Harvard-Smithsonian Center for Astrophysics Science Education Department

60 Garden Street, MS-71
Cambridge, MA 02138
Phone: 617-496-7598
Fax: 617-496-5405
Email: pstepler@cfa.harvard.edu
PARTICIPANT Q AND A

Misconception versus Misreading

- It might be hard to distinguish if students are choosing an answer because it’s a misconception or because it’s a fragment of information that they’ve learned, and they’re not reading the question, so it’s not the right answer to that question. • Participant

- Often the misconception is something that is correct, it just doesn’t apply in that particular situation. • Phillip Sadler

- So is that a matter of not reading the question well enough or is it a misconception? • Participant

- That’s why you have to have a lot of items for a concept, in hopes that you’ve hit on the understanding enough. • Phillip Sadler

Field Testing

- The field test for middle school life science is 130 items being administered to 20,000 students. We’ll select down from that number of items for the final. • Phillip Sadler

- Do they get mailed to the teachers and they mail it back? • Participant

- We’ve done it both ways. We’ve mailed them and now we have an online system, so MSPs can talk to us and we can set up an online test for you. • Phillip Sadler

- Can teachers look up the answers? • Participant

- No. This is nice because we can independently measure teacher performance on these tests. It’s all secure, so nobody knows who did what other than us, and we never report on individual MSPs, we aggregate the data. • Phillip Sadler

Difficulty vs. Discrimination

- On the Difficulty vs. Discrimination graph (page 8), the X axis is the difficulty (fraction correct), so if it were zero, that would mean nobody got it right, correct? And then if it were 1 on the Y axis, which is discrimination, it would mean that item is a perfect predictor of someone’s overall score. • Participant

- Right. • Nancy Cook Smith

Testing Teacher Pedagogical Content Knowledge

- I’m curious how you tested teacher pedagogical content knowledge (pages 9 & 10). • Participant

- It’s very simple. Using the same test given to students, we asked teachers, what is the right answer to this item and what do you predict
Types of Instruction Offered to Students

- What type of instruction were students receiving in these courses? Was it some type of special instruction? • Participant
- It was whatever the teachers normally taught. • Nancy Cook Smith
- The presumption is that if they teach in a different way they will get different results, but we looked at something quite different. We did look at what teachers did, and we didn’t find a lot of relationship between what they did in the classroom and what the gains of the students were. We found a different relationship that has never been reported in the literature. • Phillip Sadler

Link Between Identifying Misconceptions and Teacher Effectiveness

- Are we assuming that if teachers are able to identify the strong misconception, they have the teaching skills necessary for effective teaching? • Participant
- That’s a good question. Trying to get the causal nature of this out is difficult. We’re not running an experiment here, this is more an observational study in terms of looking at what happens in classrooms of teachers with different amounts of knowledge. • Phillip Sadler
- But that’s a big part of PCK right? • Participant
- It seems clear that if they don’t know the misconception they have very little chance of defeating it, but we’re not testing an assumption, we’re testing a hypothesis. • Nancy Cook Smith
- We’re not testing the teachers’ classrooms before and teachers’ classrooms after professional development programs, we are looking at these unusual aspects of teacher knowledge and trying to tease out what perhaps should be taught in professional development programs. • Phillip Sadler

Use of Misconceptions on State and National Tests

- Does NAEP or any state test use misconceptions? Are there any publicly available tests? • Participant
- NAEP pays more attention to this issue, but not exclusively like we do. And I don’t know of any state test that is based on any kind of cognitive theory that came about later than the 1950s. • Phillip Sadler
- This is a unique test development effort. • Nancy Cook Smith
Use of Misconceptions on State and National Tests

- Are the questions that you use to test students in this study available for teachers to use? • Participant

  Yes. We have two sets. Go to the MOSART Web site, where you will have to register and take a little tutorial about how these tests were constructed, and then you’ll have access to this whole battery of tests in the publicly available set. You can use those items any way you want. If you want to use them as an entire test you can, or you can select from them. The second set is a secure test and you have to negotiate with us regarding access to that. • Phillip Sadler

Addressing Misconceptions at Undergraduate and K-12 Levels

- Isn’t it fairly common at the undergraduate level to explicitly address misconceptions, especially in introductory-level courses? I’m also wondering if any of the institutes you’ve worked with are doing this at the middle or high school level, where the teacher would out and out ask, “Which of these answers would you pick?” • Participant

  I would say that there is a movement nationally in introductory college courses in physics, chemistry, and biology to address students’ misconceptions, but I would put it down below the three percent level. So it’s a movement, and there are a lot of professors involved, but if you walked into a random university you by no means would see anything like that. You might see clickers being used, but not with teachers using items that necessarily have strong misconceptions, or being able to interpret them in any way. • Phillip Sadler

  • Have you seen it at all at the K-12 level?
  • Participant

  • There are institutes that concentrate more on misconceptions. They show the videos that we produce, there are classroom discussions, teachers may actually interview other people. We’ve done institutes where we send teachers off into Harvard Square to interview people during lunch. So it’s possible to do that. • Phillip Sadler