Openness and Measurement: Two Principles for Improving Educational Practice and Shared Instructional Products

Anne K. Morris
University of Delaware

James Hiebert
University of Delaware

Two studies were conducted to identify the conditions under which instructors teaching the same mathematics teacher preparation course would continuously improve their shared instructional products (lesson plans for class sessions) using small amounts of data on preservice teacher performance. Findings indicated that when lesson-level student performance data were simply collected, by course section, the instructors could make important changes to the lessons but did not often do so. However, when the instructors were encouraged to compare data across semesters, they generated hypotheses that guided instructional improvements, which then were tested through multiple cycles. The cycles of hypothesis testing helped instructors clarify the goals for improvement, use the performance data to test whether changes were actually improvements, and reduce their tolerance for marginal student performance.

Key words: Clicker technology; Continuous improvement; Improvement science

For more than a decade, mathematics educators at our university have been working on a system of data-based continuous improvement for preparing K–8 mathematics teachers (Hiebert & Morris, 2009). For us, this means improving incrementally, each semester, the effectiveness of the four courses we offer for preparing preservice teachers to teach mathematics well to K–8 students. In particular, this means continuously improving the shared lesson plans—*the shared instructional products*—used by all instructors of the courses (Morris & Hiebert, 2011). Although there is evidence that these efforts have produced increased preservice teacher learning (Austin, 2012; Berk & Hiebert, 2009), we also have become painfully aware of the challenges of developing a sustaining system of improvement based on student performance data. To develop such a system we must learn how to help instructors use small amounts of systematically collected data to make and test important changes to the lessons.

Why do we think using small amounts of data to revise the lessons is the key to a sustaining system of improvements to the course lessons? We can identify two reasons. First, although controlled experiments have been carried out to test the effects of some changes to the lessons (e.g., Berk, Taber, Gorowara, & Poetzl, 2009), these experiments are expensive in time and effort. We cannot depend on these expensive studies to sustain a continuing improvement process. Collecting and analyzing small amounts of data is more likely to fit into the weekly work lives of instructors making the activity sustainable while also providing just enough data to test whether changes to lessons are promising improvements. Second, the model we have developed and employed to initiate the system of course improvement has been successful; but over time, we have noticed one missing ingredient—the regular collection of small amounts of data to guide proposed changes to lessons.

The model that leads to continuously improving instructional products (Morris & Hiebert, 2011) has three key features: (1) working toward the same learning goals, (2) eliciting ideas for improvement from multiple sources, and (3) using small amounts of data to repeatedly guide and test changes. In our setting, the learning goals are shared across instructors and across time because they are specified in the lesson plans used to teach each class period during the semester. Ideas are elicited from multiple sources by searching the literature for approaches to teaching preservice teachers and by soliciting ideas from new instructors (faculty and doctoral students) from year to year. But we have noticed that changes often are made to the lessons based on a rational analysis of the lesson or by anecdotes shared by instructors drawn from the responses of a few vocal students during the class session. The changes prompted by these analyses often seem to be minor changes that do not substantially improve the effectiveness of the course. How can we move instructors

Preparation of this article was supported by the National Science Foundation (Grant #0083429 to the Mid-Atlantic Center for Teaching and Learning Mathematics). The opinions expressed in the article are those of the authors and not necessarily those of the Foundation.

to use data on student performance instead of their own logic and anecdotes to improve the lessons?

Two Promising Principles: Openness and Measurement

Under the term *Improvement Science*, the quality improvement leaders and practitioners in clinical medicine and business have been studying for some time how to use small amounts of data to improve services and products (Berwick, 2003, 2005, 2008; Gawande, 2007; Kenney, 2008; Langley et al., 2009; Rother, 2009). We borrowed two principles from these fields that have led to incremental but striking improvements in the quality of outcomes: openness and measurement.

Openness means allowing one's professional practice to be open for inspection. Openness shifts professional practice from a private practice and an indicator of personal competence to a public opportunity for learning. Not just any kind of openness is productive. Learning seems to occur as data about practices and outcomes are compared and studied to identify those that most effectively achieve the desired goals.

The positive effects of openness are dependent on sharing reliably measured information on practices along with the outcomes of those practices. Comparing and studying practices requires measuring key features and outcomes of practices, not just sharing stories or anecdotes. Deciding what to measure and how to share measurements are important considerations in all fields (Kenney, 2008; Langley et al., 2009). As will be seen, we found they also are critical in the study and improvement of preservice teacher education.

Openness and Measurement in Education

Although education is different from medicine and business in several important ways (e.g., the desired goals are more difficult to specify, the practices and desired outcomes are more difficult to measure), there are lessons for education to be drawn from the two principles. Openness could facilitate improvement by breaking down the boundaries between classrooms, an impediment to improvement noted years ago (Lortie, 1975). Openness has been shown to promote the collaboration among teachers frequently called for by educators studying the essential conditions for improving classroom teaching (Gallimore & Ermeling, 2010; Little, 1990; Vescio, Ross, & Adams, 2008). Measurement of small tests of small changes that provide just enough data to make instructional decisions has been used in education for some time under the label "formative assessment" (Black & Wiliam, 1998; McManus, 2008; Wiliam, Lee, Harrison, & Black, 2004). Formative assessment can improve teachers' ability to make informed, data-based decisions. Although similar to formative assessment in some ways, the difference in the measurement process we report here is that the student learning data are used, not just to inform individual teachers' instructional decisions, but to inform the refinement of instructional products, products (like lesson plans) that hold knowledge for teaching and can be shared and passed along to new generations of instructors. The two principles—openness and measurement—were used by the first author to conduct two studies that investigated the conditions under which small amounts of data are used by instructors to make important changes to the course lessons.

The Context in Which Openness and Measurement Were Studied

Openness

Because of the culture that has been created in the mathematics K–8 teacher preparation program, a degree of openness already exists within the mathematics education group at our university. Multiple sections of three mathematics content courses are offered each semester. All instructors (faculty and doctoral students) of the same course use the same highly detailed lesson plans (see Hiebert & Morris (2009) and **Morris (2012)** for examples). Over time, instructors have become comfortable openly sharing their teaching experiences during the instructor weekly meetings. Because the same lesson plans are used, instructors are able to examine possible reasons for similar and different experiences. In the studies reported here, we were searching for the conditions of openness that would support the improvement of shared instructional products through the consistent (rather than the occasional) sharing and comparing of small amounts of data on preservice teachers' performance.

Measurement

Two kinds of measures have been used in our program to motivate and guide improvements to course lesson plans. As noted earlier, controlled experiments have been conducted, but these studies are too expensive in terms of time and effort to sustain a continuous improving system. Quiz and test scores are shared among instructors, but these measurements are too far removed from the daily lessons and not sufficiently detailed to prompt insights into problems that could be fixed or to create the moti-

vation needed for instructors to voluntarily go back and improve previous lessons. We have not yet identified the conditions that prompt instructors to consistently improve lessons based on systematically collected small amounts of lesson-relevant data.

Needed: Sharing Immediate Feedback of Preservice Teacher Performance

The following two studies describe an effort to measure preservice teachers' learning from individual lessons in a form that can be openly shared and discussed among the group of instructors to stimulate ideas for improving the lessons. With *clicker* technology now available, it is possible to create assessment items for this purpose.

Lesson-level clicker items offer two advantages: They are easy to administer, and they provide immediate feedback on how well preservice teachers in each section of a course understood a key aspect of the day's learning goal(s). This makes the data available for weekly meetings, where instructors focus on the most recently taught lessons. The items can be tightly connected to specific learning goals, making it easier for instructors to link preservice teachers' performance to particular instructional activities and teacher actions. These advantages could enable instructors to change the lesson based on preservice teachers' learning rather than on instructors' logic or anecdotes from a few vocal preservice teachers' responses.

What effect would using clicker items have on the nature of instructors' weekly discussions? Would data-based improvements be made to the lessons? The studies we describe in this article were designed to answer these questions. Because of space limitations, we will summarize the methods and results of Study 1 and present more details of Study 2.

Study 1: How Do Lesson-Level Assessment Data Affect Instructors' Efforts to Improve Lessons?

The overall question addressed in Study 1 was whether lesson-level data alone would drive the improvement process. Would simply having data about the effectiveness of lessons just taught affect the choices instructors make about how to improve the lessons? In particular, would the data prompt the instructors to engage in the kind of cause-effect reasoning that connects instructional moves to student outcomes (Gallimore, Ermeling, Saunders, & Goldenberg, 2009)? These general questions translated into three research questions:

1. Did the clicker data change the basis on which instructors revised lessons, from logic and anecdotes to student performance data?

2. What, exactly, did the instructors do with the data?

3. If instructors had access to the clicker data, were the changes they made to the lessons major (likely to affect student learning) or minor (less likely to affect student learning)?

As will be described, we found instructors could use the clicker data to develop hypotheses about major changes that might improve the lessons, but they did not do this frequently. Why?

Methods

All three instructors for the first mathematics content course for preservice K–8 teachers constituted the sample in Study 1. The course coordinator was a clinical faculty member (taught two sections); the other two instructors were doctoral students (taught one section each). As normal, the instructor group met weekly to discuss past and future lessons. At the beginning of the semester, the course coordinator sent the following instructions to the other two instructors, which set a typical agenda for the weekly meetings:

Before each meeting, please do the following so you are prepared for the discussion: (1) make a list of things that could be changed/improved for the 2 lessons we taught this week; (2) read through the 2 lessons to be taught next week; and (3) make a list of any questions you have about the lesson plans to be taught next week.

The first author wrote 25 clicker items (1–3 per lesson) to assess Lessons 7–19 (of 26).

Lessons 1–6 and Lessons 20–26 were not assessed with clicker items. This design provided an opportunity to compare the nature of the instructors' suggestions for improving the lessons before, while, and after using lesson-level clicker items. Items were given at the beginning of the lesson following the lesson that the items were assessing (e.g., Lesson 7 items were given at the beginning of Lesson 8). The clicker items were created to link tightly to the learning goals of each lesson. Here is an example.

Lesson 13 learning goal:

Given three numbers (e.g., 0.4, 6, 15) that can be related in a multiplication number sentence, preservice teachers will (a) arrange them into two valid equations (e.g., $0.4 \times 15 = ?$ and $15 \times 0.4 = ?$), (b) write the meaning of the equations (e.g., "find 4 groups of one tenth of 15," "find four one tenths of 15," "find four tenths of 15," "find 15 groups of 0.4"), (c) make [concrete] models of the equations, and (d) use the models and the meaning of the equations to help write a story situation that can be solved with the equation.

Clicker item for this goal:

A teacher asked her class to write a multiplication story problem that involved the numbers .7, 3, and 2.1 and could be solved with an equation of the form $a \times b = ?$. Four students wrote the following stories. Which of the stories are correct?

a. Brenda has 3 pounds of bananas. Sue has seven tenths more than Brenda. How many pounds of bananas does Sue have?

b. Bill has seven tenths of a pound of apples. Tom has 3 times more than Bill. How many pounds of apples does Tom have?

c. Howie has seven tenths of a pound of bananas. Stephen has 3 times as much as Howie. How many pounds of bananas does Stephen have?

d. Hilda has 3 pounds of apples. Jill has seven tenths as much as Hilda. How many pounds of apples does Jill have?

e. Both c and d are correct. (the correct answer)

The first author observed all instructor meetings, took detailed notes on the discussions of the instructors around the three agenda items, and collected the course coordinator's notes, which listed each of the proposed improvements. The first author has a collegial relationship with the instructors but did not supervise them.

Results and Discussion

Did the clicker data change the basis on which instructors revised lessons, from logic and anecdotes to student performance data? Instructors' discussions during the weekly meetings were coded to identify the rationales for the changes that were made. Three kinds of rationales were observed: (a) instructors' logic based on a rational analysis of the lesson, (b) anecdotes of individual students' responses during the lesson, and (c) clicker item data. For example, during the weekly meeting one instructor might say, "Sue asked me why we multiply to get the answer to this problem, so let's add an activity to make this clear." This was coded as an anecdote of individual students' comments. A second coder coded 20% of the data (the data for 15 of the 75 changes that the instructors decided to make over the course of the entire semester). Intercoder reliability was 93%.

On the first six lessons (not assessed with clicker items), the instructors suggested 16 changes to the lessons. The changes were motivated by 16 codeable rationales. The 16 rationales were about evenly split between instructor logic and anecdotes. On the next 13 lessons (which were assessed with clicker items), the instructors suggested 40 changes and formulated 43 rationales for those changes (a single change can be motivated by more than one rationale). Of the 43 rationales, 35 (81%) were based on instructor logic and anecdotes, and eight (19%) were based on the clicker data. On the final seven lessons (not assessed with clicker items), the instructors suggested 19 changes with 20 rationales. The basis for the rationales returned to pre-clicker patterns—the rationales were about evenly split between instructor logic and anecdotes. Instructors showed they could use the clicker data to make changes to the lessons, but they did so for only a small fraction of the lessons. What did they do with the data?

The answer to this second research question was found by coding what the instructors did with the data from each of the 25 clicker items into four categories that show increasing levels of analysis of the data: (1) no analysis, (2) discussed the meaning of the data (instructors tried to understand the nature of the student errors on the item), (3) connected the meaning of the data to their instruction (using their interpretation of student errors, instructors hypothesized instructional causes for these errors), and (4) made changes to the lesson based on the hypotheses. A second coder coded 40% of the data (the data for 10 of the 25 clicker items). Intercoder reliability was 97%.

Results showed that for 14 of the 25 items there was no analysis, for four items at least one instructor discussed the meaning of the data, for seven items the instructors worked together to connect the meaning of the data to hypotheses about a deficiency in the lesson that might explain the students' errors, and for these same seven items they made changes to the lesson based on these hypotheses. We will call these seven items *hypothesis items*. A hypothesis item is defined as an item that prompts instructors to formulate hypotheses about instructional causes for student performance on the item and to make changes to the lessons based on these hypotheses. Again, instructors showed they *could* engage in creating lesson

improvements based on a small set of relevant data, but they did not do this frequently.

Of the 25 items, the average performance across all sections for 13 items was below 70% and for 12 items the average performance was above 70%. All seven hypothesis items that prompted data-based changes were in the lower scoring set of items. In fact, 5 of the 7 hypothesis items were 5 of the 6 lowest scoring items. Apparently, instructors' attention was drawn to items on which the preservice teachers performed especially poorly.

Were the changes instructors made to the lessons major or minor? The changes instructors made during the clicker item lessons were coded into major changes (involved a significant change in the nature of an instructional activity, or the addition of a new activity) and minor changes (involved a small change in sequencing of activities, a slightly different explanation, or a small reduction in the number of problems). An example of a major change is adding a new activity to develop preservice teachers' understanding of the distributive property. Major changes are more likely than minor changes to affect preservice teachers' learning. For the major vs. minor code, a second coder coded 20% of the data. Reliability was 100%. Results suggested that basing changes on data led to major changes (5 of the 7 changes based on the seven hypothesis clicker items described above) whereas basing changes on instructor logic or anecdotes led to minor changes (30 of 33 changes based on instructor logic and anecdotes).

To summarize the results from Study 1, instructors showed they could use small amounts of data to make major changes to the lesson by forming hypotheses about what kind of change would improve preservice teachers' learning. But when just provided with the clicker items, they did not engage in this process very often. They did not connect the data to their instruction for 18 of the 25 items. When they did engage in this process, they selected lessons where the clicker data showed especially low performance. Moreover, for the 40 changes instructors made when clicker data were available, 32 of them were minor changes.

These results prompted a second study that tried to answer the following questions raised by these results. How can instructors be helped to engage in analyses of the clicker data more often, and even for items in which performance exceeds 70% correct? How can instructors be supported to use data analyses to engage in hypothesis-driven reasoning? How can instructors be encouraged to make major rather than minor changes to the lessons?

Because the results of Study 1 generated not only these questions but also some tentative answers, the second study was designed to address these questions by testing these answers. First, it seemed that instructors could be helped to engage in analyses of clicker data more often (first question posed above) if they were asked to share their clicker data for all items (including those showing high performance) during each weekly meeting. During Study 1, instructors chose to share clicker item scores for only 13 of the possible 100 scores (4 sections times 25 clicker items). In addition, to ensure that instructors evaluated all lessons, even those with clicker data above 70%, they could be asked, prior to each weekly meeting, to independently rate the effectiveness of each lesson after teaching it and decide whether improvements could be made. Continuous improvement requires constantly searching for ways to improve practices, even those that might appear to be *good enough*.

Second, we reasoned that engaging in data analyses could lead to more hypothesis-driven reasoning (second question above) if instructors could be supported in making their hypotheses explicit and making predictions about how preservice teachers' performance would improve on the items after lesson changes were implemented. During Study 1, when instructors engaged in data analysis, they looked at the poor performance of the preservice teachers, guessed what kind of change might improve this performance, changed the lesson accordingly, and left it to the Spring semester instructors to see if performance improved. Although they formed a hypothesis, they could not test their hypothesis. This makes it difficult to engage in the deeper cause-effect reasoning about instructional effects on learning that can produce *testable predictions*.

Finally, it appeared that major, rather than minor, changes could be facilitated (third question above) by pointing instructors to ideas or principles that could guide their search for substantive instructional changes that would likely make a difference for students' learning. Instructors could, for example, be asked to evaluate every lesson for whether the nature of the learning opportunities was consistent with the learning principles that drove the creation of the lessons.

Study 2: What Features Facilitate a Data-Based Sustaining Improvement System?

Guided by the questions raised by Study 1 (conducted during the Fall semester) and our tentative explanations or answers to these questions, Study 2 (conducted during the following Spring semester) was designed to change the conditions from Study 1 to test these explanations. Three

requests of the instructors changed the conditions from Study 1. First, instructors were asked to use the first few weekly meetings of the semester to look back at the clicker data from the previous semester, compare data across sections for each item, and make changes to the lessons when their analyses of the data suggested changes were needed. This was called the *Course Improvement Project*. The instructors were given an *Activity Change Report* form (**Appendix A** shows the form and the instructor's reports) to carry out the project. The aim was to engage instructors in using the clicker data from the previous semester in hopes of stimulating the formation of data-based hypotheses about changes to the lessons that could be tested during the current semester, thereby enabling the formation of predictions. Comparing Fall data to Spring data would allow instructors to create *forward-looking* hypotheses—hypotheses that included predictions that instructors could test during the semester.

The second request asked instructors to individually fill out an *Instructor Booklet* in which they entered the percentages of students in their section(s) who gave each multiple-choice response to the clicker item and then wrote a response to the following questions: "Based on the data, do you think there is a problem with the lesson that needs to be fixed? If you responded 'yes,' do you have a suggestion about how it can be fixed?" The aim of the first question was to encourage instructors to think about all lessons, even those in which performance was above 70%. The aim of the second question was to encourage instructors to formulate cause-effect hypotheses that connected the clicker data to their instruction. The booklets also contained Effectiveness Score sheets. These sheets asked instructors to give each taught lesson an Effectiveness Score (see **Appendix B**). The Effectiveness Score was the sum of the instructor's ratings on the overall effectiveness of the lesson plus the lesson's effectiveness in engaging the preservice teachers in dimensions of the two learning principles underlying the lesson design: opportunities for preservice teachers to (1) engage in productive struggle with the important mathematical ideas and (2) to construct explicit understandings of the key conceptual relationships that would support students' efforts to achieve the learning goal. The focus on the two pedagogical principles was intended to remind instructors of where they could find ideas that would produce major, rather than minor, changes to the lesson.

The third request asked instructors to share the clicker data and the Effectiveness Scores during the weekly meetings. The process for sharing the data was developed by the course coordinator: She wrote on the board the percentages of students who had correctly responded to each of the clicker items administered that week. She provided this information for seven sections individually—the four Fall sections (from Study 1) and the three Spring sections.

Methods

Sample. Two instructors who had participated in Study 1, one of the doctoral students and the clinical faculty member, participated in Study 2 and taught the three sections offered that semester. The clinical faculty member served as the course coordinator and taught two sections of the course; the doctoral student taught one section.

Methods. The 25 clicker items from Fall were used in Spring, along with one new clicker item written by the course coordinator. The first author observed all of the weekly instructor meetings and took notes on instructors' discussions of the following: clicker data, Effectiveness Scores, proposed revisions to the instruction, rationales for revisions, and the Course Improvement Project. Additional sources of data included the course coordinator's notes, the Instructor Booklets, and the Activity Change Reports from the Course Improvement Project (Appendix A). The same codes used in Study 1 were used in Study 2.

Results and Discussion

Did new conditions promote data-based improvements of lessons? An initial question is whether the change in conditions prompted a change in the way the clicker item data were used to stimulate changes to the lessons. The simple answer is yes, but with some critical caveats.

Formulating hypotheses and testing predictions increased instructors' use of data. As in Study 1, only Lessons 7–19 were assessed using clicker items. For changes suggested to these 13 lessons, 38% of the instructors' rationales for making the changes were based on clicker item results. Even for changes to the first six lessons of the course (which were not assessed with clicker items), 27% of the rationales appealed to students' performance on clicker items (while teaching lessons 7–19, instructors went back and made changes to the first six lessons based on the clicker results to these later lessons). Recall that corresponding percentages for Study 1 were 19% and 0%, respectively.

The most plausible explanation for instructors' increased use of data in Study 2 comes from the effects of the Course Improvement Project. Recall that during the first weeks of the semester, instructors were asked to review the Fall semester clicker data. As they looked across the Fall sections, they noticed a pattern of low performance in some lessons. They concluded the clicker items of interest all required a competency not addressed in the course (i.e.,

preservice teachers did not "flexibly interpret diagrams, including K–8 students' diagrams"). The instructors designed several instructional activities to develop the competency, implemented them in the Spring, and compared the Fall and Spring data to test several hypotheses (completed in several cycles described below): The new instructional activities would develop the competency, the set of clicker items all required the competency and performance on all the items would therefore improve, and the competency would transfer to related tasks and clicker items. The instructors then used the identified set of clicker items to test the changes they had made to accomplish their goal for improvement. At this point in the semester, the only request that had been made of the instructors was to fill out the Activity Change Report form.

Generating these hypotheses about what might have caused preservice teachers' low performance shifted the instructors' reasoning to a forward direction; they now were making changes to lessons, predicting their effects, and using and comparing the Fall data (collected before the changes were implemented) and the Spring data (collected after the changes were implemented) to test their predictions. This concept of forward reasoning now allows us to contrast the kind of reasoning instructors were faced with in Study 1: reasoning backward from the clicker data about what might have caused difficulties for preservice teachers. We will call this analysis *single-item retrospective analysis*—wondering what went wrong in an individual lesson without the benefit of any predictions about the outcomes. It is difficult to interpret data if explanatory hypotheses and predictions have not been made prior to collecting the data. It still is possible to formulate hypotheses based on retrospective analysis, but the hypotheses cannot be immediately tested. We propose that forward-looking hypothesis formulation and the testable predictions it generates motivates increased use of performance data.

Cycles of hypothesis testing prompted unpacking the goal for improvement. Once the instructors generated hypotheses about what changes to instruction might facilitate preservice teachers' flexibly interpreting diagrams, they carried out their hypothesis formulation, predictions, and testing in three cycles. Each cycle involved hypothesizing and making a change in the lessons, predicting the effects of the change, implementing the change and testing the effects by comparing the Fall data with the Spring data as the Spring data became available, studying the results, and revising for another cycle. The first cycle was conducted in the context of addition and subtraction, the second in the context of multiplication, and the third in the context of division. In each cycle, instructors were testing hypotheses about how their revised lessons might help preservice teachers interpret diagrams more flexibly.

Through engaging in cycles of hypothesis testing, the instructors identified several factors that appeared to affect students' ability to flexibly interpret diagrams: understanding that some number sentences can be interpreted and modeled in multiple ways, understanding the distributive property, understanding that diagrams are ambiguous if the units of measure are not explicitly identified, and understanding that quantities need to be labeled. This analysis prompted them to make further changes to the lessons after each of the first two cycles, the cycle focused on addition and subtraction and the one focused on multiplication.

By the third cycle, the instructors hypothesized that increased flexibility with diagrams would transfer to the operation of division without adding new instructional activities specifically about division. They compared the Fall and Spring data for clicker items 23 and 26 to test this hypothesis. The following discussion at an instructor meeting after the third cycle shows their subsequent attempts to *unpack the competencies* that affect students' ability to flexibly interpret diagrams in the context of division, to refine the clicker items based on this analysis, and, consequently, to understand students' thinking and difficulties at a deeper level.

Instructor A: Clicker question 23 (see Appendix A) showed [the preservice teachers] were not able to transfer some of the ideas about diagramming that we discussed for addition, subtraction, and multiplication [through the new assignments]—that is, being flexible about what the diagram represents. In Question 23, the diagram could represent either meaning of division [repeated subtraction or partitioning] and they didn't see that. They tended to see partitioning, but it could also be representing repeated subtraction. They might not be strong enough with both meanings of division so they only see one or the other. Being flexible with "what does this diagram represent?" did not transfer, but I think it's more that they aren't strong enough with division.

Instructor B: This clicker item is more about the meaning of division. They're still muddling the two interpretations together. Maybe it's too early to test for transfer.

Instructor A: So instead, maybe we could have a clicker item [at this point in the course] that tests their ability to transfer to division the idea of not having measuring units provided.

The choices would not be focused on the meanings of division, but would ask them "which number sentences could this diagram represent?" This would test their ability to transfer the idea that we have to label the measuring units or the diagram is ambiguous. Then a little later, we would have an item where the options focus on the two meanings of division, when they have a stronger understanding of the two meanings of division. Also, the students need to know that when there's not enough information, you need to be more flexible. I looked at Exam 2 problem 1 for another example of student diagrams involving division to see if they do better when it's more constrained; when they get more information on the diagram, do they do better? From the problem, I found they do okay when more information is included.

Instructor B: So maybe they don't do as well when diagrams are missing information and when they can be interpreted in multiple ways.

This conversation shows a deepening analysis of the learning goal—the instructors' target for improvement—into constituent parts. At the beginning of the third cycle, they expected "flexibility" to transfer to the operation of division. Based on multiple cycles of testing, they now hypothesized that the ability to flexibly interpret diagrams requires some concepts that cut across the four operations (e.g., the necessity of specifying units) and some concepts specific to the operation (e.g., the two meanings of division), and flexible interpretation can only be expected after those subconcepts are sufficiently robust. The instructors began to design clicker measures that would directly test their hypotheses about these specific competencies. The items would assess just one skill, and the data would then have more precise implications for instruction.

In an effort to gather even more data to inform their analysis of preservice teachers' thinking about these issues, the instructors, without prompting from the authors, looked at responses to particular exam questions. Within the context of testing their increasingly refined and forward-looking hypotheses, these exam questions no longer seemed distant and unhelpful for improvement but rather added useful information to the instructors' hypotheses for improvement. An important parenthetical point here is that the analysis of exam responses, within the context of cause-effect hypotheses of teaching-learning, is an activity that can be engaged by all teacher educators (and classroom teachers). Special clicker items are not required.

Stepping back, we believe a key lesson to draw from the cycles of hypothesis testing is that the decomposition of an initial, often vaguely defined learning goal into its component parts is a critical outcome of hypothesis formulation and refinement during sequential cycles of testing.

Unless teachers can unpack often generally stated learning goals into more precise component parts, they cannot design appropriate instruction, they cannot focus their improvement efforts in productive ways, and they cannot measure precisely enough students' performance to know if their changes are actual improvements. Unpacking learning goals is especially difficult for teachers (Morris, Hiebert, & Spitzer, 2009). Consequently, the fact that instructors in this study engaged in this unpacking process as they were driven to test their hypotheses through repeated cycles of testing instructional changes suggests this as a productive setting in which teacher educators, and classroom teachers, could decompose learning goals in a useful way.

Hypothesis testing reduced instructors' tolerance for poor performance. If instructors actually tolerate students' performance until it reaches some low threshold level, this obviously would interfere with efforts to constantly improve all the lessons. The first question in the Instructor Booklets measured this tolerance (recall that the Instructor Booklets asked instructors to individually enter the clicker data for each lesson for their section(s) and then decide whether the lesson needed improvement). As found in Study 1, when just looking at performance data, instructors in Study 2 indicated that lessons did not need improvement until clicker item performance fell below 70%. The question is: How can instructors motivate themselves to improve lessons in which the performance data rises above 70%? The answer suggested by Study 2 is that instructors should formulate forward-looking hypotheses about how to improve the lessons and engage in cycles of testing.

Evidence for the claim that hypothesis formulation and testing reduced instructors' tolerance for marginal performance can be found in the instructors' report on their Course Improvement Project (Appendix A). After their first cycle of testing focused on addition and subtraction, the instructors reported the results for clicker item 10, an item used to test the effects of a new instructional activity. The percentages of correct responses for the three sections in the Spring (after their intervention was implemented) were 82%, 76%, and 87%. Although all of the percentages exceeded the instructors' prior 70% cut-off for tolerating students' marginal performance, they pushed for more improvements in their Activity Change Report written near the end of the semester (see italicized portion of Appendix A, item 9 under Question 10).

More systematically analyzed data that support the claim that hypothesis testing reduced instructors' tolerance for marginal performance are found in Table 1. Table 1 includes the Spring data for all clicker items, divided into hypothesis and nonhypothesis items. Recall that hypothesis items are items that prompted instructors to make hypotheses about a deficiency in the lesson that might explain the students' errors and to make changes to the lesson based on these hypotheses. By comparing the Fall and Spring data, the nine hypothesis items shown in Table 1 measured the effects of the changes made in the Fall (Study 1) or in the Spring Course Improvement Project. *Nonhypothesis items* are clicker items not used to test hypotheses.

Data in the first column of the table show the preservice teachers' results for each clicker item. Data in the second column are from the Instructor Booklets. They are the instructors' individual responses to the question "Based on the data [their students' performance on the specific clicker item], do you think there is a problem with the lesson that needs to be fixed?" If the entry in the second column says "no," it means both instructors independently responded "no" to the question. If the entry says "yes," it means both instructors independently responded "yes" to the question. If the entry says "yes; no," it means Instructor A responded "yes" and Instructor B responded "no." To detect patterns, items are ordered, within nonhypothesis and hypothesis categories, with respect to instructors' belief that a lesson needed to be fixed. The shaded rows in the table highlight the clicker items that prompted at least one instructor to believe a lesson needed to be fixed. Comparing columns 1 and 2 provides a measure of the instructors' tolerance for marginal performance. Notice that when they were assessing the effectiveness of a lesson on their own, their tolerance was similar to that in Study 1—they did not believe lessons needed improving if students' performance in their section(s) was greater than 70% correct. This is apparent by the fact that, with one exception, items with at least one score below 70% are exactly those items in the shaded portions of the table.

Columns 3, 4, and 5 of Table 1 show how the instructors analyzed the data for each item during the weekly meetings as the Spring data became available and they could compare Fall and Spring data. The Spring data allowed the instructors to test their hypotheses about changes they had made to the lessons, both those made in the Fall and those made in the Spring Course Improvement Project (would the change improve student performance?). Changes to the lessons shown in column 5 are changes instructors made in the Spring after they had compared Fall and Spring data.

For the 12 nonhypothesis items where both instructors independently indicated in their Instructor Booklets that the lesson did not need to be fixed, the instructors analyzed the meaning of students' wrong answers for only 3 of the 12 items (column 3) and made no changes to the lessons based on the Spring data (column 5). For nonhypothesis items where one or both instructors independently indicated that the lesson needed to be fixed because of poor student results, the instructors analyzed the meaning of students' wrong answers at a slightly higher rate (for 2 of 5 items versus 3 of 12 items) but were not likely to change the lessons (this occurred for only 1 of the 5 items and the revision was a minor change). These results are similar to Study 1 and show that instructors often tolerate their students' performance *if* it rises above 70% and no hypotheses are driving improvement efforts.

As shown in columns 3, 4, and 5 of Table 1, treatment of the hypothesis items was different. Discussions of hypothesis items usually included analyzing the meaning of students' wrong answers *and* explicitly connecting the data to the instruction (7 of 9 hypothesis items versus 3 of 17 nonhypothesis items). The reason for the high rate for hypothesis items might be obvious but it is critical: Unlike retrospective analysis, the instructors' hypotheses identify an instructional cause and they interpret student responses within that framework. Rather than moving backward from an effect (the student data) to the identification of a cause (something in their instruction) as they must do in retrospective analysis, the hypothesis stipulates the cause (the changes they made to the instruction), and instructors use the data to test whether their hypotheses are correct. As noted earlier, this forward-looking analysis appears to encourage the use of data.

Table 1
Instructors' Analyses of the Spring Clicker Data for Hypothesis and Nonhypothesis Items

Percentage of correct responses on each clicker item in each of the three sections (Item number)	From instructor booklet: Based on the clicker data from their sections, did the instructors believe the lesson needed to be fixed?	Spring weekly meetings: Did instructors analyze the meaning of students' wrong answers?	Spring weekly meetings: Did instructors connect the data to instruction?	Spring weekly meetings: Did instructors decide to make additional changes to the lesson after they analyzed the data? If so, was it a major or minor change?
Nonhypothesis Clicker Items				
95, 97, 94 (Item 7)	No	No	No	No
91, 97, 94 (Item 14)	No	No	No	No
90, 88, 89 (Item 11)	No	No	No	No
95, 86, 83 (Item 12)	No	No	No	No
78, 89, 94 (Item 15)	No	No	No	No
81, 82, 94 (Item 8)	No	No	No	No
81, 76, 89 (Item 5)	No	No	No	No
74, 85, 72 (Item 1)	No	No	No	No
68, 73, 79 (Item 3)	No	No	No	No
74, 73, 76 (Item 24)	No	Yes	No	No
100, 100, 100 (Item 22)	No	Yes	Yes	No
91, 87, 88 (Item 2)	No	Yes	Yes	No
83, 53, 94 (Item 13)	Yes; no	Yes	Yes	Minor
48, 87, 82 (Item 17)	Yes; no	Yes	No	No
62, 64, 56 (Item 4)	Yes	No	No	No
76, 58, 56 (Item 6)	Yes	No	No	No
64, 59, 56 (Item 9)	Yes	No	No	No

Table 1—Continued

Percentage of correct responses on each clicker item in each of the three sections (Item number)	From instructor booklet: Based on the clicker data from their sections, did the instructors believe the lesson needed to be fixed?	Spring weekly meetings: Did instructors analyze the meaning of students' wrong answers?	Spring weekly meetings: Did instructors connect the data to instruction?	Spring weekly meetings: Did instructors decide to make additional changes to the lesson after they analyzed the data? If so, was it a major or minor change?
Hypothesis Clicker Items				
77, 73, 100 (Item 19)	No	No	No	No
77, 76, 94 (Item 20)	No	No	No	No
76, 87, 82 (Item 10)	No	Yes	Yes	No
86, 85, 100 (Item 18)	No	Yes	Yes	No
42, 75, 73 (Item 26)	Yes; no	Yes	Yes	Major
41, 67, 59 (Item 16)	Yes	Yes	Yes	Major
48, 46, 59 (Item 21)	Yes	Yes	Yes	Major
43, 61, 61 (Item 23)	Yes	Yes	Yes	Major
56, 60, 61 (Item 25)	Yes	Yes	Yes	Major

There was also evidence that the instructors were less tolerant of marginal performance in the context of lower-scoring hypothesis items versus lower-scoring nonhypothesis items (items in the shaded portions of Table 1). These two sets of items were treated very differently. Instructors analyzed the meaning of students' wrong answers and connected the data to the instruction (columns 3 and 4) for all five lower-scoring hypothesis items but carried out this kind of analysis for only one of the five lower-scoring nonhypothesis items. The analyses of the five lower-scoring hypothesis items all resulted in major revisions as opposed to one minor revision for the five lower-scoring nonhypothesis items (column 5).

To summarize, the results of Study 2 suggest that the kinds of data comparisons that motivate data-based improvement efforts, regardless of performance levels, are comparisons of performance before-instructional-change to after-instructional-change involving a hypothesis test. Forward-looking hypothesis testing seemed to motivate improvement efforts and decrease instructors' tolerance for marginal performance.

Non-pooled data comparing instructors' results did not promote data-based improvements. Recall that a goal of these studies was to search for conditions of open sharing of data that promote continuous improvement. A common approach to motivating U.S. teachers to improve their performance is to compare teachers in terms of their students' performance (Cohen, 1996). Indeed, debates currently are raging about how to conduct such comparisons in equitable ways (e.g., Harris, 2009). The important question for us was whether the conventional method of comparing instructors in terms of their students' achievement worked for jointly improving instructional products. How did instructors respond to the course coordinator posting clicker results, by section, for the Spring semester, at the beginning of each weekly meeting?

One way to answer the question is to check what happened when the performance differences were large, say about 25 percentage points or more. If comparing data across instructors is a condition that motivates them to examine the data and search for improvements, these big differences in performance should prompt discussions among instructors that would explore the reasons for the better performance. There were five such items. For one of the items (Item 19), Instructor B's section scored 23% and 27% higher than Instructor A's two sections respectively. But the instructors carried out no analysis of the item. Instructor A did not ask Instructor B what about her instruction might have produced the higher performance.

For the remaining four items, one of Instructor A's sections (not always the same section) scored noticeably lower than the other two sections. When the data were shared for two of the four items (Items 16 and 26), the instructors did not refer to the performance differences across sections and the data did not prompt them to compare their instruction in the three sections. For the remaining two items (Items 13 and 17), the instructors made some attempt to explain away the data; they attributed the results to some aspect of the students' behavior (e.g., lack of studying) rather than the instruction. For example, for Item 17, the following exchange took place.

Instructor A: [Puts the data on the board: 82% correct (Instructor B's section), 48% correct (Instructor A's first section), 87% correct (Instructor A's second section)]

Instructor B: [Laughing] What's with the section? Maybe you do better the second time around?

Instructor A: Thirty-three percent said D was the answer. I asked, "Why is B not correct?" They just said, "Oh. They weren't careful."

Thus for all five items showing a large performance difference, the open sharing of data that compared the performance of the instructors' students did not prompt the instructors to compare the details of their instruction in order to identify potentially better practices.

Because comparing student performance data across instructors is often assumed to create conditions that motivate instructors to improve, we consider why this did not happen, even for these large differences. When instructors are implementing the same lessons using similar instructional practices, student performance measures can show one instructor or class is doing better, but the instructors cannot explain why. If the differences were caused by variants of the intended instruction, the instructors could not reconstruct these details and consequently had no basis on which to hypothesize causes for the differences. So, in some cases they did what many instructors do—they attributed the results to students' behavior. This is a problem for a continuous improvement process because it means that instructors, at least momentarily, are relinquishing responsibility for the effects of their instruction and removing themselves from the improvement process (Kenney, 2008).

A second problem with the instructors' comparing section-by-section data (four sections from the Fall and three sections from the Spring) was that the data were too difficult to interpret in this form; the instructors could not always tell whether the instructional changes they were making in the Spring were improvements. For example, in the instructors' report of the Course Improvement Project (Appendix A), the instructors compared the percentages of correct responses in the seven sections and made the following statement.

> **For clicker items 23 and 26, the instructors thought that the data showed slight improvement, but we were not satisfied. We thought that [preservice teachers] would be able to transfer the types of knowledge we discussed with interpreting diagrams for addition, subtraction, and multiplication to division diagrams. However, the data showed that this is not really the case.**

The results from the two items were treated the same ("the data showed that this is not really the case [that transfer would occur]"). However, if the data from all classes in a given semester are pooled, a comparison reveals students performed significantly better in the Spring (after the instructional changes were implemented) than they did in the Fall (before the changes were implemented) on item 23 ($z = 3.47$, $p < .001$) whereas there were no significant differences between the Fall and the Spring for item 26. This type of comparison would have provided information the instructors needed when they were attempting to unpack the competencies that contribute to students' ability to flexibly interpret diagrams (described above). They could have analyzed the competencies that are involved in answering item 23 versus item 26 and used that analysis to understand the effects of their new instruction.

These two issues indicate that the section-by-section sharing of data was problematic. What kinds of comparisons would be more productive? We can offer the following suggestions.

Results from Study 2 suggest that before- and after-instructional-change comparisons involving a hypothesis test changed the acceptable level of failure and motivated data-based improvement efforts. So, comparisons that support hypothesis testing appear to be more productive. The findings above suggest this happens when data are compared across time but not across instructors or classrooms. On a practical level, this means pooling the data across classes for each clicker item and comparing the data across semesters.

Comparing data across time supports hypothesis testing *when* instructors can distinguish improvements from just changes. One way to enable these distinctions is to statistically analyze the pooled data across semesters. These analyses can determine whether a change to the instruction was a real improvement. Because some changes produce important but small improvements, several hypothesis-testing cycles might be needed to accomplish an improvement goal. A test of significance can be used to compare the beginning overall percentage of correct responses (before any changes are made) with the final overall percentage of correct responses (after a number of related revisions have been made to address the same goal).

A consequence of statistically testing pooled performance data across time is that it places attention where it belongs—on the instructional products (e.g., lesson plans) rather than on individual instructors. A shared instructional product is designed to minimize across-instructor variability so the instruction can be linked to student outcomes (Morris & Hiebert, 2011). Comparing data across instructors highlights this variability and focuses attention back onto the individual instructors. A before- and after-instructional-change comparison, with data pooled across instructors, returns the focus to the instructional product.

Before concluding this section, we need to add an important caveat. Past experience has taught us that, under some conditions, comparing results across instructors can be productive. If instructors are discussing relative successes and failures of recent lessons and find that one instructor had more success than others, this discussion can encourage the successful instructor to share her approach with other instructors often saying, "Oh, that's a good idea. I'll try that," usually to good effect. The difference in the studies we are reporting here is that the clicker *data* created differences among instructors in preservice teachers' performance with no hypotheses to explain the differences. This left the instructors with no productive avenue of analysis.

What effects did the Effectiveness Scores have on the instructors' improvement efforts? Effectiveness Scores had two effects on improvement efforts. First, they drew attention to instructional problems. The instructors were more likely to think there were problems with the instruction that needed to be fixed when they were assigning Effectiveness Scores than when they were evaluating the clicker data. Instructor A and Instructor B believed that 38% and 27%, respectively, of the lessons with clicker items needed to be fixed, based on the clicker data. In contrast, Instructors A and Instructor B gave Effectiveness Scores indicating there was some deficiency in the lesson to 69% and 54%, respectively, of the lessons assessed by clicker items.

Second, the Effectiveness Scores prompted major instructional changes rather than just minor changes. This was most pronounced for the productive student struggle dimensions of the Effectiveness Score. Instructor A and Instructor B assigned non-perfect scores to one or both of the student struggle dimensions for 33% and 37% of all course lessons respectively. Productive student struggle then became a major focus of the discussions during the instructor meetings. Whereas two changes in the Fall semester involved increasing productive student struggle (both major changes), 13 changes of this type were made in the Spring (11 were major changes). In contrast to the tendency of the instructors in the Fall to make minor changes (see Study 1), the instructors in the Spring made a number of major changes by developing cognitively demanding activities to increase the level of productive student struggle.

Based on these two results, it appears that an especially useful role for Effectiveness Scores is to point instructors to fundamental features of lessons that could be considered for improvement. On a daily basis, as instructors assign each lesson a score, they are reminded of learning principles that drive the lessons and could become the target for cycles of specific testable major changes.

In this study, the Effectiveness Scores focused the instructors' attention on productive student struggle. Typical of fundamental learning goals in education, productive struggle is quite vague. Without a more refined and elaborated definition, how do teachers improve lessons with respect to this learning principle? The results of Study 2 reveal that one significant contribution of the repeated cycles of hypothesis testing can be clarifying the meaning of these principles and enabling more focused improvement efforts.

In future semesters, other foundational principles could be selected. In this way, teacher educators can customize Effectiveness Scores to shine the spotlight on learning principles on which they would like to focus their improvement process.

Does collecting and providing lesson-level data to instructors produce increased learning? The ultimate question is whether all of this work to improve the quality of the instructional products increases preservice teachers' learning. To answer this question, clicker data from all sections in the Fall and the Spring, respectively, were pooled for each of the 25 items that were administered in both Fall and Spring and the percentages of correct responses were compared across semesters. The comparisons test the effects of the lesson changes that were made before administering the Spring clicker items. Z-scores were calculated and a 99% confidence level was used as the standard for increased learning. Percentages of correct responses did not significantly improve for any of the 16 nonhypothesis items (see Table 1 for these items). For the hypothesis items, performance increased significantly for 5 of the 9 items (Items 10, 16, 18, 21, and 23). For one of the remaining four items (Item 25) only minor changes were made and consequently performance increases would not be as likely. For the final three items (Items 19, 20, and 26), major changes were made but no significant performance increases were found on these items. This might be due to the difficulty of the mathematical topics being studied; several cycles of improvement work might be needed.

Implications for Teacher Education

The promise of the principles of openness and measurement to make productive use of small amounts of data was realized, but it was the details of applications that mattered. Given increased attention to data-driven instruction, in both K–12 classrooms and teacher education programs, the details become critical. Taken together, Studies 1 and 2 suggest that just providing instructors data on students' performance is not enough to prompt frequent and continuing improvements in instruction. Data on past performance, even at the lesson level, do not, by themselves suggest what changes should be made. Perhaps because the data do not reveal immediate implications for instruction, they do not seem to motivate instructors to engage in the time-consuming work of analyzing instruction and learning in a cause-effect way and making improvements.

What seems to be critical is to provide instructors (who are teaching toward the same learning goals) with guidelines of how to *use* the performance data. In particular, the results of these studies suggest that productive work is facilitated by asking instructors to use small amounts of targeted performance data to formulate hypotheses about what changes to instruction might improve performance, to make these changes and predict what effect they will have, and then to measure the effects of these changes. The results can then prompt a new cycle of hypothesis development, changes to instruction with predicted effects, and measurement. These hypothesis-testing cycles seemed to be the key to instructors' success in making major improvements to the lessons of interest.

One consequence of these testing cycles appeared to be a decrease in instructors' tolerance of preservice teachers' marginal performance. Apparently, instructors' views can change from "that performance is good enough" to the view that marginal performance is something that can be studied further and improved even more ("Try again. Fail again. Fail better" (Beckett, 1983).).

A second consequence of hypothesis testing cycles appeared to be a clarification of the learning goals into more specific subgoals or parts that can be studied with more precision. A somewhat unique feature of education compared with other fields is that many learning goals are quite general. Helping students make progress in achieving them requires breaking them into constituent parts and studying how to help students achieve more precisely stated subgoals.

We began by borrowing from traditions of improvement in other fields, especially medicine and business, to hypothesize that the principles of openness and measurement could play an important role in improving teacher education, and teaching more generally. We now look back to this tradition of improvement and note that the cycles of hypothesis testing we described bear a striking resemblance to something called PDSA (Plan, Do, Study, Adapt) cycles in Improvement Science (Langley et al., 2009; Nolan, Schall, Berwick, & Roessner, 1996).

The PDSA cycle is an iterative process in which improvements are developed, tried, studied, and then refined, over and over, until quality is improved. The first step is to *plan* the test. The plan involves creating a potential solution based on a clear hypothesis that generates an explicit, testable prediction about the outcomes. The second step is to *do* or carry out the test—that is, to execute the plan and assess the results of it by collecting small amounts of relevant data. Next comes *studying* the result and comparing the result to the prediction. It is this comparison that produces learning. Finally, based on the results of the analysis, a decision is made on how to act. What next step is warranted based on what was learned? Should the change be adopted, adapted and tested further, or abandoned? We believe that the improvement process in teaching could benefit from learning more about the processes that Improvement Science has to offer.

Finally, we believe the results indicate that the *method* of sharing and comparing data matters. Many educators have too quickly accepted the assumption that comparing teachers on their students' performance will lead to improved teaching. As we have described, other methods

of sharing and comparing these data might be more productive. For example, comparing data across time seems to work better than comparing data across instructors.

Our aim in these two studies was to learn something about the conditions under which the open sharing of small amounts of data can, and will, be used by instructors to improve their instructional products. We hope the suggestions we offered will be tested further, both in teacher education and classroom settings.

References

Austin, J. (2012). *Continuous improvement of mathematics teacher education.* Presentation at the annual meeting of the National Council of Teachers of Mathematics. Philadelphia, PA.

Beckett, S. (1983). *Worstward ho.* New York: Grove Press.

Berk, D., & Hiebert, J. (2009). Improving the mathematics preparation of elementary teachers one lesson at a time. *Teachers and Teaching—Theory and Practice, 15*(3), 337–356.

Berk, D., Taber, S. B., Gorowara, C. C., & Poetzl, C. (2009). Developing prospective elementary teachers' flexibility in the domain of proportional reasoning. *Mathematical Thinking and Learning, 11*(3), 113–135.

Berwick, D. M. (2003). Errors today and errors tomorrow. *New England Journal of Medicine, 348*(25), 2570–2572.

Berwick, D. M. (2005, December 12). Keys to safer hospitals: A set of simple precautions could prevent 100,000 needless deaths every year. *Newsweek, 146*(24), 76–78.

Berwick, D. M. (2008). The science of improvement. *Journal of the American Medical Association, 299*(10), 1182–1184.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in education, 5*(1), 7–74.

Cohen, D. K. (1996). Rewarding teachers for student performance. In S. H. Fuhrman & J. A. O'Day (Eds.), *Reward and reform* (pp. 60–112). San Francisco, CA: Jossey-Bass.

Gallimore, R., & Ermeling, B. A. (2010). Five keys to effective teacher learning teams. *Education Week, 29*(29).

Gallimore, R., Ermeling, B. A., Saunders, W. M., & Goldenberg, C. (2009). Moving the learning of teaching closer to practice: Teacher education implications of school-based inquiry teams. *Elementary School Journal, 109*, 537–553.

Gawande, A. (2007). *Better: A surgeon's notes on performance.* New York, NY: Metropolitan Books/ Henry Holt.

Harris, D. N. (2009). Would accountability based on teacher value added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*, 319–350.

Hiebert, J., & Morris, A. K. (2009). Building a knowledge base for teacher education: An experience in K–8 mathematics teacher education. *Elementary School Journal, 109*(5), 475–490.

Kenney, C. (2008). *The best practice: How the new quality movement is transforming medicine.* New York, NY: PublicAffairs.

Langley, G. J., Moen, R., Nolan, K. M., Nolan, T. W., Norman, C. L., & Provost, L. P. (2009). *The improvement guide: A practical approach to enhancing organizational performance* (2nd ed). San Francisco, CA: Wiley.

Little, J. W. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record, 91*, 509–536.

Lortie, D. (1975). *Schoolteacher: A Sociological Study.* Chicago, IL: University of Chicago Press.

McManus, S. (2008). *Attributes of effective formative assessment.* Washington, DC: Council of Chief State School Officers.

Morris, A. K. (2012). Using "lack of fidelity" to improve teaching. *Mathematics Teacher Educator, 1*(1), 71–101.

Morris, A. K., & Hiebert, J. (2011). Creating shared instructional products: An alternative approach to improving teaching. *Educational Researcher, 40*(1), 5–14.

Morris, A. K., Hiebert, J., & Spitzer, S. M. (2009). Mathematical knowledge for teaching in planning and evaluating instruction: What can pre-service teachers learn? *Journal for Research in Mathematics Education, 40*(5), 491–529.

Nolan, T. W., Schall, M. W., Berwick, D. M., & Roessner, J. (1996). *Reducing delays and waiting times throughout the healthcare system.* Boston, MA: IHI.

Rother, M. (2009). *Toyota Kata: Managing people for improvement, adaptiveness, and superior results.* New York, NY: McGraw-Hill.

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education, 24*, 80–91.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*(1), 49–65.

Authors

Anne K. Morris, School of Education, 105B Willard Hall, University of Delaware, Newark DE 19716; abmorris@udel.edu

James Hiebert, School of Education, 107A Willard Hall,
University of Delaware, Newark DE 19716;
hiebert@udel.edu

Appendix A: Study 2: Activity Change Report Forms That Were Completed by the Instructors to Report on Their Course Improvement Project

Activity Change Report

An instructor group should fill out this form whenever they change an activity in the lesson plans in response to the student results from a clicker item. **Please fill out one form for each activity that is changed.** In some cases, a clicker item or items might suggest several activities should be changed. For example, students' performance on a number of items might suggest that the lessons are not effective in developing pre-service teachers' ability to write story problems. The instructor group might then develop a rationale (see #5 in the form) to address this issue. For example, they might identify a new way of teaching the writing of story problems that they believe will be more effective. This rationale—that pre-service teachers are not performing well on story problem clicker items and that this new approach to the teaching of writing story problems will be more effective—has implications for several lessons (e.g., lessons on writing story problems for addition and subtraction, lessons on writing story problems for multiplication and division). The instructor group might then change several activities across several lessons to address this rationale. If multiple activities are changed because of the instructor group's rationale, then the separate forms for each activity should be stapled together. This will allow subsequent instructor groups to see all of the changes that were made to address that rationale. In other words, if a common rationale underlies a group of changes, please staple the separate forms together; activity change report forms should be grouped by rationale.

Instructors' Activity Change Report I

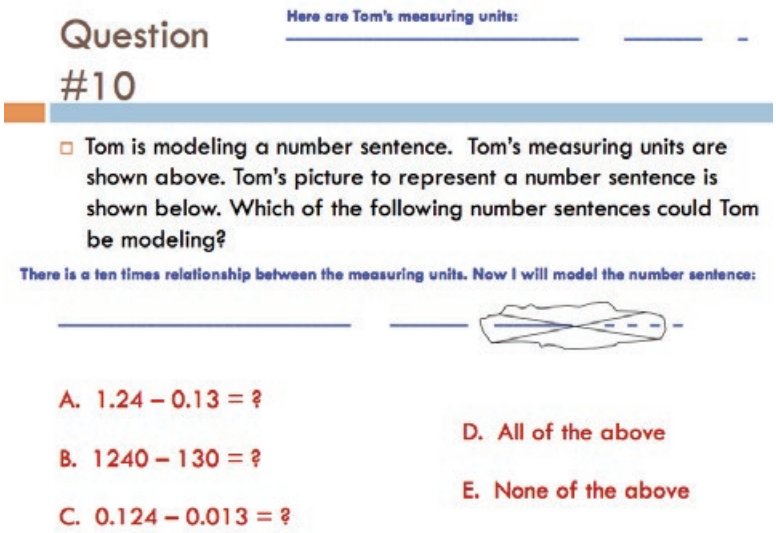
1. Names of the instructors: . . .

2. Semester and Year: . . .

3. Lesson Number, Lesson Name, and Activity Number where the change was made:

   Lesson 10, AddSubMeanII, Extra Homework

4. Please list the clicker item(s) that prompted the reported change to the lesson activity. Write the item(s) here, along with the multiple-choice answers.



5. What was the instructor group's rationale for making the change to the lesson activity? Please explain how the rationale is connected to the clicker items listed in #4.

   Pre-service teachers struggle with interpreting student work when diagrams are missing labels, measuring units, or are otherwise unclear. We thought that they needed to improve on their flexibility in how they interpret diagrams when information is missing. Moreover, we wanted PSTs to realize the importance of having all of this information on a diagram to make the diagram clearly linked to one number sentence.

6. Please cut and paste the old version of the activity here:

   A previous version of this activity did not exist.

7. Please cut and paste the new version of the activity here:

   See attached document. [Here is a sample problem from the instructors' new instructional activities:]

   Consider the following diagram that Troy (a fourth grader) drew below. Does the diagram clearly illustrate a *specific* number sentence? If yes, state the number sentence. If not, explain why not.

8. Please provide the student results from each clicker item listed in #4 before and after you made the change(s) to the lesson activity.

Percent of students responding correctly:	Clicker Item 10
Before you made the change (i.e., last semester):	
Section 1	**38%**
Section 2	**23%**
Section 3	**5%**
Section 4	**71%**
After you made the change (i.e., this semester):	
Section 1	**82%**
Section 2	**76%**
Section 3	**87%**

9. Does the instructor group think the change was successful? Are the changes in clicker performance enough for the instructors to be satisfied or do they think there still is work to be done?

> ***The instructors feel that the clicker item data provide evidence that there was a huge improvement in PSTs' ability to flexibly interpret diagrams for addition and subtraction of decimal numbers. Some improvement could be made to determine what specifically is the most difficult piece of interpreting diagrams with missing parts (i.e., is it the measuring units missing that makes it hard, is it the missing labels, or is it something else?). We think that we could be more proactive in linking a specific homework problem with a specific clicker item to measure more specifically what the PSTs have trouble with.***

Instructors' Activity Change Report II

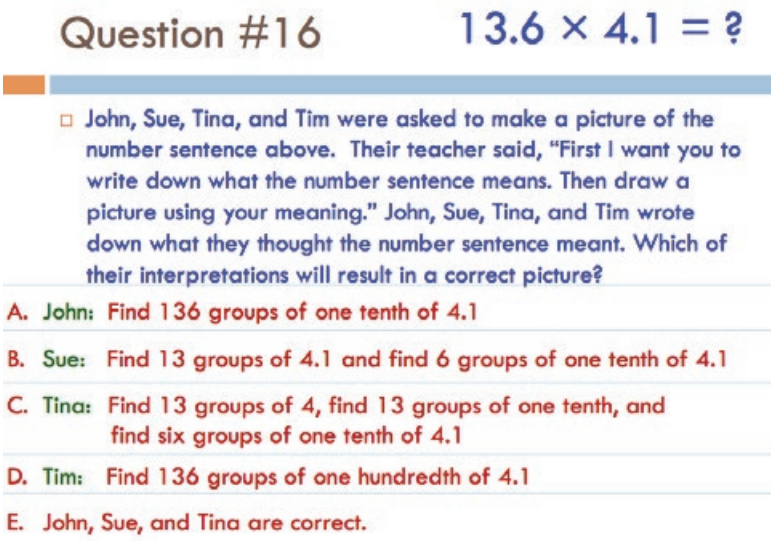
1. Names of the instructors: . . .

2. Semester and Year: . . .

3. Lesson Number, Lesson Name, and Activity Number where the change was made:

Lesson 14, MultMeaningI, Activity 2, Lesson 15, MultMeaningI, Extra Homework

4. Please list the clicker item(s) that prompted the reported change to the lesson activity. Write the item(s) here, along with the multiple-choice answers.



Question #19

- If you are solving 0.23 x 4.1 with a diagram on graph paper, you would get the right answer if you found the area of:

A. 23 groups of one-hundredth of 4.1

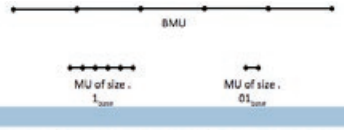
B. 2 groups of one-tenth of 4.1 and 3 groups of one-hundredth of 4.1

C. 2 groups of one-tenth of 4 and 2 groups of one-tenth of 0.1 and 3 groups of one-hundredth of 4 and 3 groups of one-hundredth of 0.1

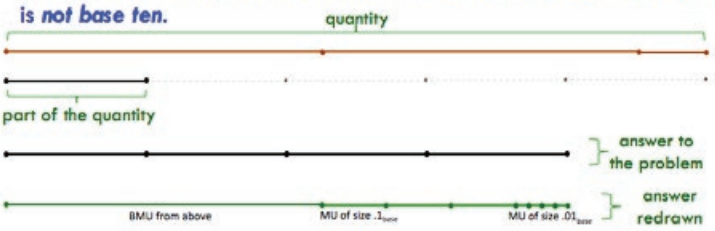
D. a, b, and c are all correct

E. Only a and b are correct

Question #20



- Using the measuring units as shown above, Michelle has drawn the picture below to represent and solve a number sentence in a base that is ***not base ten.***



What is the number sentence she is solving, and what is the answer?

A. $0.4_{fifteen} \times 2.1_{fifteen} = ?$ and the answer is $1.34_{fifteen}$

B. $0.4_{five} \times 2.1_{five} = ?$ and the answer is 1.34_{five}

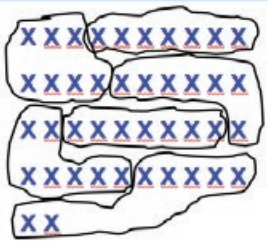
C. $4_{five} \times 2.1_{five} = ?$ and the answer is 1.34_{five}

D. None above

E. Both b and c could be correct

Question #23

- Ted is modeling a division number sentence at the right. What number sentence could he be modeling?



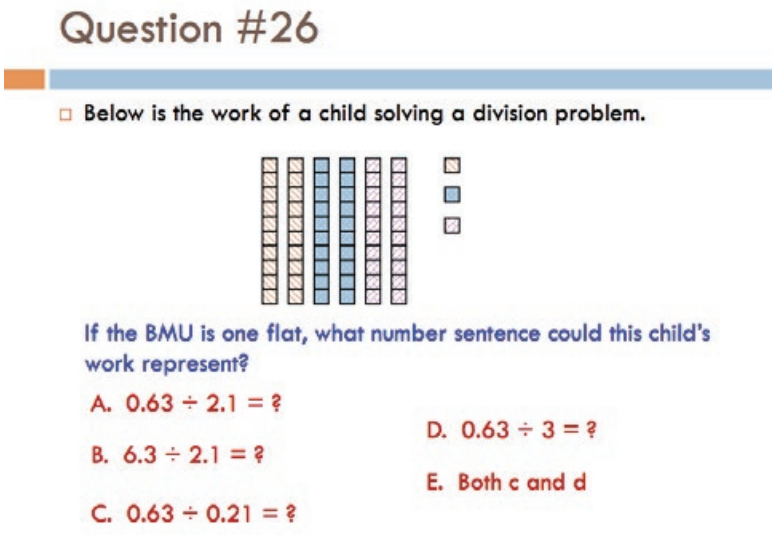
A. Ted could be modeling .42 ÷ .07 = ? using the repeated subtraction interpretation of division.

B. Ted could be modeling .42 ÷ 6 = ? using the partitioning interpretation of division.

C. Ted could be modeling 4.2 ÷ 6 = ? using the repeated subtraction interpretation of division.

D. None of the above

E. Both a and b



5. What was the instructor group's rationale for making the change to the lesson activity? Please explain how the rationale is connected to the clicker items listed in #4.

   Pre-service teachers struggle with interpreting student work when diagrams are missing labels, measuring units, or are otherwise unclear. We thought that they needed to improve on their flexibility in how they interpret diagrams when information is missing. Moreover, we wanted PSTs to realize the importance of having all of this information on a diagram to make the diagram clearly linked to one number sentence. We added a similar extra homework for addition and subtraction. Now, we added an extra homework for multiplication in the hopes that they would be able to transfer these ideas to division as well.

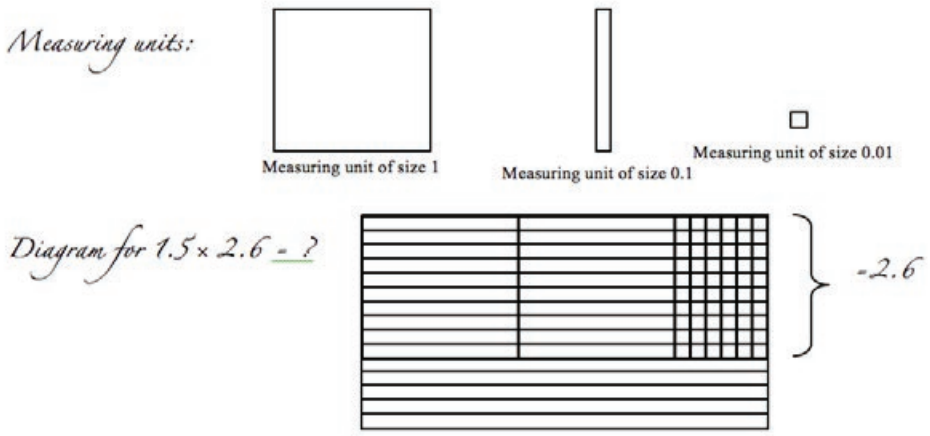
6. Please cut and paste the old version of the activity here:

   A previous version of these activities did not exist.

7. Please cut and paste the new version of the activity here:

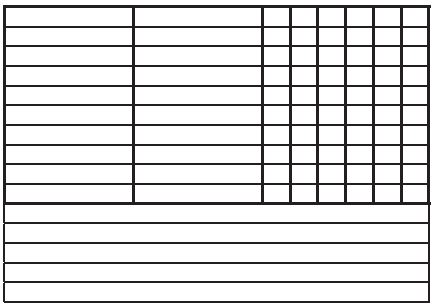
   For Activity 1, we added the following question to promote discussion of different ways to solve the multiplication problem: *What would be the all-at-once interpretation and the by-place interpretation if we broke up the 3.1 into place value parts?* For the extra homework, please see the attached document. [Here is a sample problem from the instructors' new instructional activities:]

   Consider the following diagram for 1.5 × 2.6 = ?

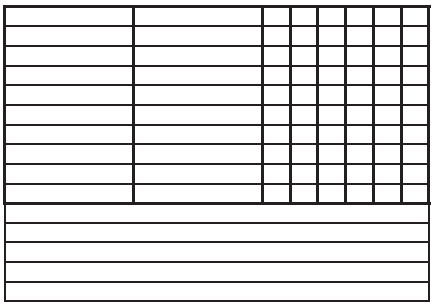


Three students say the following. Under each student's statement, show how to highlight what the student sees in the diagram.

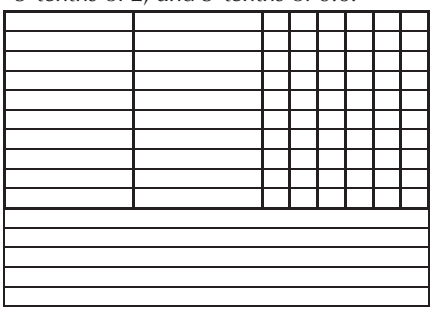
Pamela: I see 1 group of 2.6 and 5-tenths of 2.6!



Cory: I see 15-tenths of 2.6!



Sara: I see 1 group of 2, 1 group of 0.6, 5-tenths of 2, and 5-tenths of 0.6!



8. Please provide the student results from each clicker item listed in #4 before and after you made the change(s) to the lesson activity.

Percent of students responding correctly:	Clicker Item 16	Clicker Item 19	Clicker Item 20	Clicker Item 23	Clicker Item 26
Before you made the change (i.e., last semester):					
Section 1	**32%**	**76%**	**81%**	**14%**	**46%**
Section 2	**27%**	**66%**	**71%**	**33%**	**50%**
Section 3	**45%**	**52%**	**79%**	**53%**	**67%**
Section 4	**31%**	**81%**	**58%**	**24%**	**68%**
After you made the change (i.e., this semester):					
Section 1	**59%**	**100%**	**94%**	**61%**	**73%**
Section 2	**41%**	**77%**	**77%**	**43%**	**42%**
Section 3	**67%**	**73%**	**76%**	**61%**	**75%**

9. Does the instructor group think the change was successful? Are the changes in clicker performance enough for the instructors to be satisfied or do you think there still is work to be done?

The instructors feel that the clicker item data from clicker items 16 and 19 provide some evidence that there was satisfactory improvement in PSTs' ability to solve multiplication of decimal problems in more ways. We feel that the added discussion in Lesson 14 and some of the extra homework problems helped PSTs think about multiple ways to view multiplication. We still feel that improvements could be made to help PSTs understand the distributive property.

For clicker item 20, the instructors thought that the data showed slight improvement because of the extra practice that they had from the extra homework. But all of the extra practice was with an area model. We think that to improve what we did, discrete models and linear models should be included in addition to just the area model for multiplication.

For clicker items 23 and 26, the instructors thought that the data showed slight improvement, but we were not satisfied. We thought that the PSTs would be able to transfer the types of knowledge we discussed with interpreting diagrams for addition, subtraction, and multiplication to division diagrams. However, the data showed that this is not really the case. So, the instructors think it would be beneficial to have PSTs complete extra homework problems that focus on interpreting division diagrams with missing parts and where both partitioning and repeated subtraction are within the same diagram.

(Return to page 135)

Appendix B: Effectiveness Score Sheet

Please give a score to this lesson immediately after teaching it. Assign a score of 0 (low) to 2 (high) for each of the five criteria below. Indicate your score for each of the five criteria by checking the cell for 0, 1, or 2 in the table below. Then calculate the total score for the lesson by adding the five scores and write the score on the Total Score line below the table.

	Score		
	0 points (Not at all)	**1 point (Some, but needs improvement)**	**2 points (No improvement needed)**
The lesson succeeded in helping students achieve the learning goals.			
The lesson activities align well with the learning goals for the lesson.			
The lesson activities provided students with an opportunity to struggle with the critical mathematical concepts.			
The students had time to struggle with the critical mathematical concepts.			
During the lesson, the students received a clear explanation of the conceptual relationships among mathematical ideas, representations, and/or procedures from myself or other students.			

Total Score for Lesson x: _____

(Return to page 135)