Learning Together: A User-Friendly Tool to Support Research on STEM Education Interventions

March 2012

Evelyn M. Gordon
Iris R. Weiss
Joan D. Pasley
Daniel J. Heck

horizon
RESEARCH, INC.

Learning Together: A User-Friendly Tool to Support Research on STEM Education Interventions was prepared with support from the National Science Foundation under grant number EHR-0445398. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Learning Together: A User-Friendly Tool to Support Research on STEM Education Interventions

Federally funded K-12 science, technology, engineering, and mathematics (STEM) education projects are generally expected not only to use what is already known in designing and implementing interventions but also to add to the knowledge base. STEM education projects supported by NSF, the Department of Education, and other funders, often involve a mix of people with quite different backgrounds and prior experiences: STEM faculty; STEM education faculty; district supervisors; and master teachers. Involving people with diverse backgrounds helps bring a great deal of expertise to the table, but at the same time it can create communication challenges, as the terminology that some people find extremely useful can seem like just a lot of jargon to others. This document provides a brief introduction to research on K-12 STEM education interventions; it is intended to help people who may be new to social science research understand some of the key issues. We include some terminology commonly used in social science, but the emphasis is on developing concepts that project teams can refer to as they design and implement research.

This document focuses on a particular kind of social science research: understanding how and why interventions work the way they do, in particular interventions aimed at improving teaching and learning in K-12 STEM education. Although both evaluation and research have the potential to generate useful knowledge, and many of the ideas described in this document apply to both, the focus here is on research conducted in conjunction with K-12 STEM education interventions.[1] The primary distinction is the purpose of the knowledge generation. While evaluation focuses on assessing the quality and impact of project activities in a particular context, research has a broader purpose: understanding if, how, when, and why particular kinds of interventions are likely to be effective, as part of theory building.

When investigating the impact of interventions, it is essential to be clear about the focus of the research and to design studies to determine the degree to which any differences that are detected can reasonably be attributed to the intervention.[2] High-quality research depends on the use of appropriate instruments, including making sure that a study measures what the research team intended to measure (*validity*), and that the measures can be used consistently over time and by different people (*reliability*). Both research design and measurement considerations are essential; research that has serious flaws in either area will not provide useful information. Planning a study, therefore, is not a linear process; rather, research design, instrumentation, and practical issues must be weighed and revisited as a plan is developed.

Recent reviews of the empirical literature in mathematics and science education have identified substantial deficiencies in the research base as a whole, and a need for the field to adopt more rigorous standards of evidence (e.g., Heck, 2008; Hill & Shih, 2009; MSP-KMD, 2010; National

[1] For additional information about social science research methods see *The Research Methods Knowledge Base* (Trochim, W. M.), available for download from http://www.socialresearchmethods.net/kb/ For information specific to conducting project evaluations see *The 2010 user-friendly handbook for program evaluation* (Frechtling, 2010), available for download from http://www.westat.com/pdf/projects/2010ufhb.pdf.

[2] Although exploratory research can be extremely beneficial, e.g., in getting a sense of how interventions are playing out so they can be fine-tuned, and in generating hypotheses for systematic investigation, this document considers the more formal research that will be shared with the field.

Research Council, 2004). The following sections address a number of key considerations, focusing on the identified deficiencies, in designing and reporting research on STEM interventions.

Section 1. Focusing the Research

Research in any field needs to build on what is already known, with the goal of developing more complete understanding. In the case of research on K-12 STEM education interventions, the goal is to develop understanding of what works, for whom, and under what conditions, in order to help improve teaching and learning at scale. Teams must consider how their research can contribute to the existing knowledge base as they select interventions and outcomes on which to focus their research. The project's theory of action, outlining the chain of events that connects the interventions to the desired outcomes, can help the team develop research questions by identifying specific relationships that merit investigation. Key considerations for contributing to the knowledge base, selecting interventions, and using a theory of action to reflect upon the research and interventions are examined in this section.

A project may have multiple components to its intervention, any one of which might target multiple goals. However, given the expense of conducting high-quality research, and the limited resources that are typically available, projects will likely conduct research on only a subset of their efforts. For example, a project may be incorporating some components of professional development that have already been heavily studied and some components that have not. Although it is always helpful to learn more about the conditions under which a particular approach is effective, a research team may decide that it can make more of a contribution to the knowledge base by devoting research resources to the less studied activities. Another research team might make a different choice, deciding to investigate the effectiveness of a reduced version of a successful intensive intervention, or the effectiveness of an intervention when provided by a less experienced team, to generate knowledge for improving teaching and learning at scale.

Research on an intervention requires an explicit description of the intervention that distinguishes it from other, possibly similar interventions. STEM education interventions are frequently complex and interactive, and inevitably somewhat different from one implementation to the next. A study might look at an intervention that combines several features that prior research has identified as effective, considered as a whole. Alternatively, a study might focus on individual features of an intervention to improve the field's understanding of their functions. In either case, as part of investigating the impact of "it," researchers must clearly describe what "it" is, i.e., the defining elements that characterize the design and implementation of their intervention.

Education interventions are undertaken for a purpose; ultimately they are intended to change student outcomes for the better. In some cases researchers may work directly with students, but often interventions are focused on teachers (or other educators) who will subsequently work with students. In these instances, there are teacher outcomes in addition to student outcomes that are ripe for study. For both students and teachers, interventions may target any of a number of short and long-term outcomes, for example: teachers' or students' beliefs, content knowledge, or attitudes; teachers' and/or students' classroom practices; or principals' support for new teaching practices. Researchers need to determine which outcomes to study.

Prior to conducting research on an intervention, a study team needs to sketch out the project's "*theory of action*," considering in detail how the project activities are expected to lead to the desired outcomes. Being explicit about the theory of action helps the project designers to consider whether the planned interventions have a reasonable chance of producing the desired outcomes, and if not, how they might be strengthened.

In terms of research, in describing the theory of action, the project team is laying out a series of hypotheses, any one of which can be investigated. For example, a professional development program for teachers of mathematics might have the following theory of action:

- If teachers explore important mathematics ideas, they will deepen their understanding of these concepts.
- If teachers with solid content understanding have opportunities to consider how their students' instructional materials are designed to address grade-appropriate content goals in these areas, they will be better able to provide effective instruction.
- If knowledgeable and skilled teachers have a supportive context, including adequate time for instruction in mathematics, they will be more likely to implement effective instruction.
- If students experience higher quality instruction, they will achieve a better understanding of the targeted mathematics ideas.

It is important to note that a project might engage teachers in learning content that is more advanced than what their students will encounter for any of a number of purposes. If a project's goal is primarily to learn about how to effectively deepen teacher content knowledge, leaving application of that knowledge for a later time, then the theory of action is expressed in the first bullet, and research would be limited to exploring that link. In that case, the project might investigate not only the overall impact on teacher content knowledge, but also whether other key variables, e.g., teacher background, make a difference in the effectiveness of the project's interventions on teacher content knowledge. If the goals of an intervention extend to classroom applications, research might still focus on the relationship between the interventions and teacher content knowledge, or the research might also include the relationship between teacher content knowledge and one or more classroom instruction variables. Decisions about where to focus the research depend on how the interventions are expected to make a difference, as well as investigators' interests in adding to the knowledge base, and what is feasible to study based on time and resource constraints.

Sometimes it is hard to make the case that there is a plausible theory of action for a component of an intervention, however laudable the goal. For example, many years ago a project planned to "transform elementary science instruction" in a large number of schools, in part by having a scientist provide a demonstration lesson to each class once a semester. In critiquing the plans, the project evaluators noted that this aspect of the intervention was unlikely to achieve its goals; even if the demonstration lessons were top notch, there was little reason to believe that classroom teachers, who would provide the lion's share of the science instruction, would be willing and able to teach differently based on observing those lessons. Although the project team decided to go ahead and implement this component, in part to honor promises they had made to the schools, they recognized that devoting extensive resources to studying that effort was not warranted.

Section 2. Designing Studies

After developing research questions about relationships between the interventions and anticipated outcomes within the context of the theory of action comes the work of designing a research study to address those questions. At a minimum, such a study will include ensuring that the interventions occur followed by an examination of the outcomes, through measurement or description. Often, a goal of research on interventions is establishing that an intervention **caused** desirable outcomes, in which case the research design would include a comparison of the outcomes with and without the interventions. Researchers also need to evaluate which other factors that are likely to affect outcomes, in addition to the project's activities, should be included in the study. This section describes key considerations for developing a conceptual study design: identifying variables, specifying analysis strategies, and addressing potential alternative explanations.

Using the theory of action, researchers need to identify information that will enable them to answer their research questions. In general, this information will include: what participants experience when the interventions are implemented; the outcomes of interest; and additional factors the project identifies as potentially affecting the outcomes. For example, research on the impacts of a professional development intervention on classroom practice and student outcomes may need to account for the possible moderating influence on classroom practice of the instructional materials that treatment and comparison teachers are expected to use. An intervention may enhance teacher content knowledge sufficiently to result in improved classroom practice in cases where the instructional materials are well designed, but not enough so teachers can critique and revise poorly designed materials.

If earlier research has identified specific elements of an intervention that led to the particular outcomes a research team is interested in, then it will be important to collect information about those elements, along with information about contextual conditions that differ from those in earlier research. When investigating the effects of a less studied intervention, on the other hand, it may be more appropriate to document the intervention through a rich description and include a broader range of anticipated outcomes or potential mediating factors.

Researchers need also to consider when particular information should be collected, and from whom, in order to examine the relationship between interventions and outcomes. In a qualitative study of a little-researched intervention, data collection could take the form of observations and interviews with participants, focusing on descriptions of what happened and participants' interpretations of the role of the intervention in their actions. In many instances, information will be collected in order to make comparisons and detect changes. One common strategy is to compare participants and non-participants. Another possibility is to collect information from participants both before and after the interventions. Combining these two strategies can strengthen the research design by establishing that the participants and non-participants were similar to begin with and that changes were not due to something other than the interventions, such as the passage of time or increased familiarity with an instrument used to collect data. In some situations when it is not feasible to collect data prior to an intervention or from a comparable group, comparisons between participants and an external standard can provide evidence of an intervention's impact. For example, student work samples could be used to show understanding well beyond what is normally expected from students at that grade level as part of making a case that the interventions were effective.

One of the key considerations in designing research is the need to determine the extent to which any observed changes can reasonably be attributed to the interventions, or if there are other, more plausible explanations for those changes. For example, if a state added science to a high-stakes assessment program during the course of a study, it is possible that the assessment and not the interventions was responsible for differences in how much time teachers devoted to science instruction. Having a comparison group that experienced the same assessment changes but not the intervention and did not increase time spent on science would help the research team make the case that the intervention was effective. Of course, not any comparison group will do. A study may find that students of "treated" teachers did better on an assessment than students of teachers who did not attend the professional development. The intervention may well have been effective, but in the absence of evidence that the participating and non-participating teachers—and their students—were comparable to begin with, there is a good possibility that initial differences, and not the effectiveness of the intervention, were responsible for the results. Having a comparison group, collecting appropriate baseline data to show that the treatment and comparison group teachers were initially similar, and confirming that there were no other important interventions at play all support the case that differences are attributable to the study interventions.

Well-designed research can help avoid unfairly concluding that a treatment does not work. For example, if there are only a handful of teachers in the experimental and comparison groups, it would take a very large difference to be statistically significant, so an intervention that in fact results in meaningful improvements in classroom instruction might be wrongly dismissed as ineffective. Collecting quantitative data is not a good use of resources unless the study design is sufficiently powerful to detect a meaningful difference. In a situation such as this, a qualitative study that described classroom instruction and provided evidence that teachers in the treatment group had applied what they had learned during the intervention in planning or enacting lessons might pave the way for larger studies focused on specific improvements.

Often, researchers are interested in student outcomes. However, interventions are usually delivered to (or by) teachers who are assigned to treatment and comparison groups. Except in the rare case where students are randomly assigned to groups, the number of teachers in the study is much more important than the number of students. For example, if a teacher-intervention study had treatment and comparison groups with 1,200 students in each, the statistical comparisons would need to be done using the 40 or so teachers in each group. The need to use the appropriate unit of analysis—in this case teachers rather than individual students—has implications for sample size.[3] Having 300 students in each group might sound like "enough," but probably would not be sufficient since the comparison would be between 10 teachers in each group.

In some cases, researchers are particularly interested in exploring the conditions under which a particular intervention is and is not effective in achieving its goals, e.g., with teachers who have different amounts of experience, or teach at different grade levels, or in different kinds of school contexts. It is important to recognize that the more ways the data will be split apart (*disaggregated*), the larger the sample size that will be needed, with a sufficient number of participants in each group to provide meaningful results.

Finally, designing studies requires being realistic about the costs of collecting and analyzing data. One of the most common mistakes in social science research, including research on STEM education interventions, is collecting more data than the researchers can afford to analyze. For example, while videotaping classrooms of treatment and comparison groups provides an opportunity to obtain very valuable information, systematic analysis of the resulting videos can take several hours per lesson, not to mention the costs of training observers and monitoring the analysis process over time.

[3] There are statistical techniques (e.g., hierarchical linear modeling) that utilize all of the available data in nested designs (e.g., students within teachers' classes) but maintain the appropriate unit of analysis.

Section 3. Generating Empirical Evidence

Section 2 addressed considerations for developing a conceptual design for a research study, but a design is not complete without a plan for data collection and analysis, including identifying instruments that will be used to gather information about the variables of interest, and converting the raw data into a form that can be used to address the research questions. This section addresses considerations for selecting appropriate instruments and planning data collection and analysis.

In designing a study, the research team will have specified the defining characteristics of the intervention that is being investigated, whether it is an existing intervention being studied under new conditions, a modification of an existing intervention, or a newly-developed intervention. However, rarely is an intervention implemented exactly as planned—differences of interpretation among intervention providers, the starting points and needs of particular groups of participants, and time constraints, among other factors, can lead to differences in the nature of implementation. Whether through logs maintained by intervention providers, observations by members of the research team, or other means, it is important to document the key elements of implementation so a study will be able to explore the relationship between the intervention and the outcomes of interest. Similarly, it is important to document the extent to which individuals participate in the interventions in order to be able to relate extent of participation in particular components to the desired outcomes.

The availability of high-quality instruments is a key consideration in study design, especially in describing/measuring outcomes of interest; it makes little sense to attempt to understand the conditions under which an intervention is effective if a study will not be able to detect effectiveness. The problem may be a lack of existing high-quality instruments, e.g., to measure teacher knowledge of student misconceptions in a particular content area. Or the problem may be that the existing measures are very expensive to administer and score, or are otherwise not feasible for use in a given context. Before moving ahead with a study design, research teams need to make sure that there are valid, reliable, feasible instruments available for the variables of interest, both outcomes and factors in addition to the interventions that might affect outcomes.

Using or modifying existing instruments to collect information is generally preferable to developing new ones, as creating high-quality instruments can be both expensive and time consuming, and typically requires specialized knowledge. In considering instruments, it is important to keep in mind that the same term may have different meanings in different projects, and to select instruments that address the research team's meanings (construct *validity*). For example, a science education intervention that emphasizes interpretation of evidence to develop conceptual understanding might find that a content assessment that tests memorization of terms would not give participants an opportunity to reveal what they have learned.

In addition, instruments must have *sensitivity* to anticipated differences in order to be useful. That is, some of the data they are capturing must be likely to change over time (if the study involves collecting information about participants at multiple time points) or be different for treatment and comparison groups. If an assessment is too easy, participants will have high scores on it prior to the intervention, so the study would not be able to detect an impact even if the intervention is in fact effective. On the other hand, if an assessment is too difficult, the anticipated score increases might be too small to be statistically significant. Similarly, questionnaire items, interview questions, and observations must focus on outcome measures that have the potential to be affected by the intervention.

Of course, instruments that are being considered must be not only appropriate but also feasible for use in the particular context, considering burden on participants as well as the capacity and resources available for data collection and analysis. Ideally, instruments the team identifies as both appropriate and feasible for use in their situation will have been tested to provide evidence that the information they generate is trustworthy. For instruments in which the researcher using the instrument must make judgments, such as scoring open-ended assessment items or observing classroom instruction, it is important to ensure that the researchers who will be involved will make similar judgments about the same raw data (*inter-rater reliability*).

Sometimes a research team will not be able to locate existing appropriate, high-quality instruments for some of the variables of interest. What then? Should the research team omit those variables from the research design, or develop their own instruments? In most cases, it would be advisable to do the former. Developing drafting, testing, and refining instruments is difficult and time consuming; rarely does a research team have the capacity and resources to do it well. And trying to detect relationships with poor measures of some of the variables is not a fruitful pursuit.

A study design will specify sample sizes, ensuring a sufficient number of cases to answer the research questions. Assuming the research team has identified high-quality instruments that are both appropriate and feasible for use in this situation, decisions will need to be made about how much data to collect about each of the nodes in the project's theory of action, e.g., in studying a professional development intervention, how much to focus on teacher knowledge and how much on classroom practice.

STEM education research on complex phenomena such as professional development programs or teaching practice can often benefit from using multiple instruments and multiple data sources. For example, to describe a professional development intervention adequately, members of the research team might collect data on a sample of sessions using an observation protocol to help in interpreting facilitator logs that document the entire intervention. Or, a project using a state assessment of student mathematics achievement as a broad measure of student learning might also administer an assessment more tightly linked to the project's learning goals to capture changes where the sensitivity of the state assessment is likely to be limited. In other instances, getting multiple perspectives may provide a more complete understanding of an outcome than collecting data from a single group of people. A teacher's classroom practice is likely to be described differently by the teacher, his or her students, a mathematics coach working with the teacher, and the school principal; a case study of classroom practice would likely be enhanced by representing several of those viewpoints.

Section 4. Going from Plan to Action

No matter how elegant a study design, no results will be generated until the design is put into action. Moving from plan to action does not happen instantaneously, and implementation can be delayed or interrupted by both foreseeable and unexpected events. Fortunately, many logistical matters can be anticipated, planned for, and even partially addressed while final details of the study design are being refined. This section describes considerations for preparing to carry out the research, including: data collection, reduction, and analysis; ensuring that qualified intervention providers, data collectors/ preparers, and data analysts are available; and seeking appropriate permissions to conduct the study.

Earlier sections focused on identifying key variables and instruments to measure or describe those variables. The study plans will also need to include procedures for converting the raw data to forms that can be used to address the research questions. For example, consider a study of the relationship between extent of emphasis on standards progressions in professional development and particular components of classroom practice. Data on how these progressions are addressed could be collected via observation, with rich descriptions of the professional development offered at various locations; and/or interviews with professional development providers; and or interviews/questionnaires administered to participating teachers. Eventually, the observation and interview data would need to be coded or otherwise "reduced" so the analyst could look for patterns between the nature and extent of emphasis on the standards progressions and the classroom variables of interest. Similarly, teacher responses to a series of related questionnaire items might be combined into a composite score that could be analyzed in relation to one or more classroom practice variables.

Sometimes study findings are not convincing because the researchers are the ones who provided the intervention, and there is the possibility that because of their investment in the treatment they were predisposed to see changes that others likely would not have seen. Having someone who was not involved in designing or implementing the intervention collect at least some of the data can help rule out *investigator bias* as the explanation for the results. For some data sources, this principle also applies to the steps that prepare raw data for subsequent analysis, such as scoring open-ended assessment responses or coding observation data. Researchers involved in data collection and preparation will likely need training to establish a shared understanding of what to observe for and/or how to code responses; such training provides additional protection against bias as well as increasing the reliability and validity of the data collection and preparation.

Investigator bias can also be a concern for analysis of the data. As with the earlier stages, including knowledgeable but less involved researchers in some phases of data analysis, and using established principles for identifying themes in data can reduce the likelihood of such bias. Participant reviews of findings based on interviews or observations, often called member checks, can ensure that researchers have not misinterpreted participants' meanings. A deliberate search for evidence that contradicts a researcher's theories or tentative findings is a vital step in qualitative analysis to protect against investigator bias.

The analysis plan should be reassessed as data are collected and prepared for analysis, to ensure that the original plan is still appropriate. In some cases, the planned statistical analyses may not be appropriate, e.g., if the data do not cover the range of possible responses or do not have a "normal distribution." Furthermore, issues such as low response rates, participants leaving the study before completion, a great deal of missing data, or different patterns of missing data among groups of participants can influence both statistical and qualitative analysis results. Prior to other analysis, it is helpful to identify any discrepancies between the data collected and the data anticipated in the research design and to consider if there will be implications for interpreting the results.

Early results, including researchers' observations of patterns in the data, can also suggest new ideas that were not part of the original research questions. It is important to note, however, that there are likely to be limitations to how thoroughly such new ideas can be addressed in the current study. For example, a study may not have sufficient sample sizes to support additional statistical analyses; or information that emerges as relevant may not have been systematically collected. Despite these limitations, exploratory analyses can help identify directions for future studies to investigate new ideas.

While a small research study may be undertaken by the core research team itself, larger ventures may well require additional staff, either to deliver the intervention, collect data, and/or prepare data for analysis. When assessing the need for additional staff, it is important to consider not only the requisite tasks but also the timeframe for completing them. For example, one or two observers might be able to conduct 40 classroom observations over the course of a semester, but additional people would be required if 40 observations had to be conducted in a single week, e.g., if the study focused on a particular topic that was scheduled to be taught at that time in all classes at a given grade level. The qualifications and training needs of the people who will be providing the interventions and/or collecting the data must also be considered, whether they are members of the core study team or brought in for specific tasks. Similarly, if sophisticated statistical analyses are planned, it may be necessary to bring in people with the necessary expertise to carry out the specified statistical analyses.

Education research studies are typically required to undergo an Institutional Review Board (IRB) review, to ensure that participants' rights are being protected. Depending on the complexity of the study and the particular IRB, such reviews can be time consuming, so research teams need to plan ahead to ensure they have the necessary approval to begin a study on time. For example, an IRB may require the project to get written permission from schools/school districts involved in the research as well as participant consent, including parental consent for minors.

Section 5. Sharing What Has Been Learned

Research results can and should be used to improve project interventions, but research teams also have a broader responsibility to share what has been learned, contributing to the knowledge base. Once shared, the knowledge can be used by practitioners making decisions about related interventions as well as by future researchers who can incorporate what has been learned into their design process. Of course, research reports should include the findings, and make the case that the interventions were (or were not) effective in particular ways. Research reports also need to describe the research methods and interventions in sufficient detail so readers will be able to understand what was done as a basis for interpreting the results. In this section we consider the information that should be included in research reports.

The research report should include a description of the study itself, including the purpose and methods for the study. In reporting qualitative research, it is important to provide information about the researchers, in addition to the study design and data collection and analysis methods, to establish trustworthiness. Journals frequently include guidelines for describing studies in their calls for submissions, and the American Educational Research Association (2006) has published reporting standards for social science research that can provide useful guidance.

The intervention itself is an integral part of any research study on STEM education interventions, yet reviews of the research literature have found that interventions are often described in only very general terms (Sztajn, 2011; Weiss & Miller, 2009). For example, the following is a typical description of an intervention in the literature on deepening teacher mathematics/science content knowledge:

> The professional development was provided during three weeks in the summer. Participating teachers were engaged in solving a series of challenging mathematics problems, meeting periodically in grade-level groups to discuss implications for their instruction. Follow up

> sessions during the academic year provided an opportunity for teachers to share their experiences in applying what they had learned in the summer to their classroom practice and pose questions to project staff.

Even if the research was very well designed and implemented, the fact that the intervention description was so vague greatly limits what the reader can take away from the study. Suppose the intervention was shown to be effective in deepening teacher content knowledge, improving classroom practice, and in turn, student achievement; readers have learned that "it" works, but would not know very much about what "it" was, and therefore would not be able either to implement a similar intervention or to design research to test the approach in a different context. Similarly, if an intervention was found to be ineffective, or partially effective, it would be difficult for a reader to consider (or another researcher to test) whether particular modifications would improve the results. In reporting research, teams should give as much of the flavor of the interventions as possible given space limitations. For example, documentation of interventions should include:

- Who delivered various parts of the intervention, including their backgrounds, preparation, and any other relationships they had to the participants and the study;
- What strategies comprised the intervention, and in what settings the participants encountered the various strategies;
- The timing and sequence of the various parts of the intervention;
- Expectations regarding what teachers would do as a result of participation, such as making instructional changes or providing services for other teachers; and
- Other relevant professional development or support participants had received during the time frame of the study.

If space is not available for such information, reports should point the reader to a web site or other places where they can get additional information.

In addition to describing the intervention, it is important to describe both the participants and their contexts. In cases where the research examined how the interventions played out in multiple contexts, the report needs to include information about the conditions under which an intervention was and was not effective. But whether or not results are available separately for each of the different contextual factors, it is important to describe the composition of the study sample in terms of variables that might "matter." For example, if the research was conducted with very experienced high school computer science teachers in resource-rich schools, the findings might not apply to elementary teachers, or to high school teachers who have less experience and/or work in more challenging contexts. Information about the participants and their contexts is important to practitioners, for assessing whether the findings are likely to hold for their target audiences and contexts. Such information is also important to other researchers for identifying gaps in the knowledge base and designing additional studies.

In studies of interventions intended to improve student outcomes by deepening teachers' content knowledge, thorough documentation would include contextual information such as:

- Participants' content, education, and teaching backgrounds;
- Whether participation in the professional development was mandatory or voluntary;
- Participants' teaching assignments, including grade levels and courses;
- Instructional materials and support resources in use in the district, schools, and classrooms;

- The policy context, including standards and assessments that may influence teachers' instructional decisions, and requirements/other incentives for ongoing teacher professional development;
- School and district administrative support for the intervention and/or its goals for teacher learning and instructional change;
- The history of related improvement efforts in the school or district; and
- School and community demographics.

Not only will providing detailed descriptions of the interventions and study participants/contexts, help readers in interpreting the results, this kind of documentation will enable future research syntheses to more fully describe what has been learned about the kinds of interventions that are effective for achieving particular goals, for whom, and under what conditions.

At the heart of research reports are statements of the claims being made from the study, the evidence and reasoning that are being used to support those claims, along with any disconfirming or surprising results. In addition to ethical considerations, it is important to report ambiguous or contradictory evidence, or evidence that results were inconsistent across locations or individuals, to provide a fuller understanding of research results. Such evidence, especially surprising results, may suggest directions for future research efforts.

Although research on interventions should be designed to rule out the most likely alternative explanations, no study can eliminate all of them, and study reports need to be explicit about the limitations of the research they are describing. Any alternative explanations that have been ruled out, including the evidence against those explanations, should be described as part of the support for the claims being made. The remaining alternative explanations (*threats to validity*) should be clearly described, both so readers can consider them in interpreting the results, and so other researchers can develop studies that directly address those threats in their designs.

Finally, research reports provide researchers with an opportunity to identify future lines of inquiry related to their work. In some cases, new hypotheses may have emerged from a study, hypotheses that can be further refined through more direct examination, or systematically tested in future studies. Additional studies testing the extent to which the reported findings generalize, e.g., to different grade ranges, contexts, or content areas, may also be warranted. Report authors should identify what appear to be the most important or promising directions for future research based on their work.

Section 6. Getting Feedback on the Study

Sections 1 through 5 describe considerations for a research team to bear in mind when making decisions about the design and implementation of a STEM education research study. Making those decisions involves negotiating trade-offs among possibilities that offer different combinations of practical and theoretical advantages. Getting periodic feedback from people who are familiar with the study but not involved in the day-to-day decision making can be very helpful both initially and as the research progresses. This section addresses considerations for selecting people to provide feedback, identifying points when consultation with these knowledgeable outsiders is likely to be most useful.

Feedback on research studies can be provided by project evaluators and/or advisory boards if the study has the resources to support these roles. A smaller effort might be able to identify a colleague who is not directly involved with the research to serve as a "critical friend," providing periodic feedback on the study design and implementation.

In the early stages of a study, it is advisable to have the draft design reviewed by one or more individuals external to the research team. An outside reviewer can assess the merits of the design independent of the trade-offs that went into making decisions, and might suggest a need to revisit those decisions. External review will be especially important if the design includes statistical analysis beyond the expertise of team members, a complicated design involving the measurement of a large number of variables or consideration of many outside factors, and/or research approaches that are new to the research team. In addition, if the research team members are quite similar in background and approach, it may be helpful to get feedback from people whose theoretical perspectives are somewhat different. (At the same time, criticism offered by a colleague with diametrically opposing views on what is central to improving STEM education may not be constructive.)

Once a study is underway, a knowledgeable outsider can serve as a sounding board for managing data collection or analysis issues. For example, if data analysis proves more time consuming than anticipated, outside advice can provide a neutral perspective in helping the research team consider which of the many potentially useful analyses are most central to the research questions. Evaluators/advisors/critical friends may also be helpful in assessing the need for specialized knowledge for completing specific tasks and identifying experts with the necessary knowledge.

Finally, feedback in the dissemination phase of research studies can help ensure that the research products communicate clearly to people who have not been immersed in the work. In particular, external reviewers can help the research team improve manuscripts before they are submitted for publication, including (a) noting terms that need to be further defined; (b) pointing out where underlying assumptions need to be made more explicit; (c) suggesting ways to strengthen the case for claims; (d) pointing out strengths and implications of the study that should be emphasized; and (e) identifying any additional unresolved alternative explanations that should be noted. In addition, people external to the research team who have followed the progress of a study may be able to suggest additional audiences that would be interested in the results and outlets for reaching them.

Section 7. Conclusion

The "Reflection Questions" in Table 1 summarize the ideas addressed in this document. They can be used to structure the research team's discussions as a study is designed and implemented, and they can provide a useful starting point for feedback from people external to the team. It is important to note that there is no one right answer to these questions; rather they are intended to help in identifying important issues for consideration. In addition, these ideas are interrelated and the answer to one reflection question likely has implications for the answers to others.

The vignette following Table 1 illustrates that there are multiple ways to go about generating useful knowledge, even about the same intervention in the same context. Planning what to study, and how, is an iterative process, involving negotiating trade-offs from beginning to end. Our hope is that this tool is in fact "user-friendly," and that it will help study teams be explicit about the various trade-offs in planning, implementing, and reporting research to expand the field's understanding of how to improve STEM education.

Table 1: Reflection Questions

Section 1: Focusing the Research

1.1 What are we trying to contribute to the knowledge base?

   a. Knowledge about the impact of previously-studied interventions—under different conditions, on different outcomes, and/or with design modifications
   b. Knowledge about new kinds of interventions

1.2 What particular interventions are we studying?

   a. What are the defining elements of our interventions?
   b. Are we studying the intervention as a whole, or specific elements of the intervention?

1.3 What relationships between our interventions and outcomes will we investigate?

   a. What outcomes are we interested in studying?
   b. How do we anticipate our interventions leading to these outcomes?

Section 2: Designing Studies

2.1 What variables do we need to describe/measure in order to address our research questions?

   a. About interim and ultimate outcomes
   b. About other factors that are likely to affect achievement of our goals

2.2 What analyses will we need to conduct to examine the relationships between our interventions and outcomes?

   a. What kinds of differences are we trying to detect (e.g., over time, between groups)?
   b. What analytic approaches will we use to detect anticipated differences?
   c. What are the appropriate entities (e.g., teacher, school) to use for our analyses?
   d. What study sample size and composition will be needed?

2.3 How can we determine the extent to which our interventions are responsible for any differences we detect?

   a. If we plan to compare groups, how will we ensure that they are initially similar?
   b. What alternative explanations for differences can we account for in our study design?

Section 3: Generating Empirical Evidence

3.1 What information should be documented about the implementation of the various parts of our interventions at each study site?

   a. What pieces of the interventions were implemented, how, and by whom?
   b. What data should we keep on participants and their participation?

3.2 What existing instruments can we use to collect information about outcomes, and about other factors that may affect achievement of our goals, specifically instruments that:

   a. Define the important variables as we do?
   b. Have the potential to detect differences in outcomes of interest?
   c. Will provide results that we can trust?
   d. Are feasible for use in our research with respect to budget, expertise, and burden on participants?

3.3 If we cannot locate existing appropriate, high-quality, feasible instruments for measuring particular variables of interest, do we have the capacity and resources to develop and test our own instruments, or should we omit those variables from our research?

3.4 How will we ensure that the data we collect provide a sound basis for our analyses?

   a. Are we appropriately distributing data collection efforts across the range of variables we have identified as important?
   b. Are we appropriately using multiple sources of information at particularly key junctures, using complementary data collection approaches, and capturing different perspectives?

Table 1: Reflection Questions

Section 4: Going from Plan to Action

4.1 How will we ensure that data collection and analysis are conducted appropriately?

- a. Have we built in checks to reduce the likelihood of bias in collecting data and preparing them for analysis?
- b. What are our plans for preparing raw data so they are ready for analysis?
- c. How will we reduce the likelihood of bias in our data analysis?
- d. How will we screen our data to determine if our analysis plan needs to be adjusted?

4.2 Do we have all of the capacity and resources necessary for this study?

- a. For delivering the interventions
- b. For collecting, preparing, and analyzing the data

4.3 Are we appropriately documenting our protection of participants’ rights?

- a. Do we have Institutional Review Board (IRB) approval for our research?
- b. Are we obtaining the necessary permission to conduct this study (e.g., district permission, participant consent)?

Section 5: Sharing What Has Been Learned

5.1 What information should we report about our interventions, participants, and context to:

- a. Help readers interpret and judge our findings?
- b. Enable readers to consider the applicability of the findings to different contexts?
- c. Enable readers to implement/adapt our interventions for different contexts?

5.2 What information should we report about our research methods?

- a. What data did we collect to address our research questions?
- b. How did we analyze those data?

5.3 What did we learn?

- a. What claims can we make?
- b. What evidence and reasoning can we use to support our claims?
- c. What alternative explanations for our findings have we ruled out, and how?
- d. What alternative explanations for our findings have not been ruled out, and how do we acknowledge those in our reports?

5.4 What are the implications of our study for future research?

- a. What additional studies are most important in order to test the extent to which our findings apply more generally, e.g., to different grade ranges or contexts?
- b. What hypotheses emerged from our study that will need to be systematically tested?

Section 6: Getting Feedback on the Study

6.1 What kinds of feedback will we need, from whom, during the planning and implementation of this study?

6.2 When should we seek external input?

Vignette

The CONNECTIONS project provides both summer and academic-year professional development (PD) for K-12 STEM teachers in a large district. The focus is on “targeted ideas” of science and mathematics that have been identified as particularly important for students in the district to learn: these ideas are included in state content standards, addressed in student instructional materials, and highlighted in state assessments. STEM faculty involved in the project have “unpacked the content,” highlighting the connections among the targeted ideas within and across grades. After a pilot year to “get the kinks out” of the interventions, the project plans to randomly assign schools to one of two groups, an initial treatment group and a delayed treatment group that could serve as a comparison; teacher applicants will then be scheduled for professional development according to their school’s assignment.

The intervention was designed based on what is known about effective PD. Teachers are given opportunities to explore mathematics/science concepts as learners, both to deepen their content knowledge and to enable them to experience instruction that reflects current understanding of how people learn. Participants are then asked to reflect on the implications of the professional development for their classroom practice, with teachers in each grade/course focusing on the student instructional materials they are expected to use, including: (1) identifying the key learning goals of selected lessons; (2) considering what difficulties students might have in achieving the learning goals; and (3) analyzing student work from those lessons. At various junctures in the program, cross-grade groups of teachers discuss how student understanding of the key ideas is expected to develop over time, to facilitate K-12 articulation. The project also provides an orientation session for principals so they will understand what their teachers are experiencing.

The project evaluation is designed to address both the quality of the various project components and their overall impact, in both mathematics and science. Each year, the evaluators will observe sessions; collect both survey and interview data from participants; administer tests of teacher content knowledge; analyze teacher and student assessment data; and make recommendations for fine-tuning the interventions based on the evaluation findings.

The pilot year evaluation report noted that the various components of the project were generally of high quality and were very well received, with participants indicating that the PD was both engaging and helpful. On the average, teacher scores on project-developed mathematics/science content knowledge tests were substantially higher after PD than before PD. And the average gains on the district’s end-of-year mathematics assessments for students of participating teachers were higher compared to those teachers’ similar students the previous year. However, when the project team looked at the data more closely, they were surprised to see that some middle and high school teachers had large student assessment score gains in all of their classes, some had minimal gains across the board, and others had sizable gains in their “honors” or “advanced” classes but minimal gains in others. The research team wants to understand what is going on, and why, so they can ensure that all students in their project schools will benefit, and also help the broader field anticipate barriers they might encounter and design interventions accordingly.

In thinking about what might explain the findings, the research team considers what is already known about influencing factors, and identifies several variables as potentially important: teacher pedagogical content knowledge; teacher beliefs about student learning based on perceived student ability level; and school-based support.

The team discusses exploring the following hypotheses for their results:

- Teachers implemented the new teaching strategies in all of their classes, but some teachers were better at differentiating instruction to meet students’ needs.

- Teachers used different ones of the new strategies in their regular and advanced classes.
- Some teachers implemented new strategies more in their advanced classes, leading to greater gains for those classes.

As they begin to flesh out a possible study, they consider focusing on classroom practice to see if the teaching strategies promoted in the PD are evident in their program's "graduates," and what differences, if any, there are in teachers' instruction in different level classes. They quickly realize that they have a depth versus breadth decision on their hands: survey all of the teachers or do intensive data collection with a subset of the teachers? They are leaning towards studying a small number of teachers fairly intensively, including observing classrooms and interviewing teachers to get a more in-depth understanding about what is happening in classrooms, and why.

The team also discusses whether to conduct their research on the entire K-12 spectrum in both mathematics and science, or to focus on a particular grade or grade range. Consistent with the idea of depth over breadth, they talk about restricting their research to a single grade, and to a subset of the content addressed in the professional development.

The team identifies 7th grade mathematics as a potential focus for several reasons. First, there are high-quality measures of teacher content knowledge available for teachers at that level, including one which aligns well with the project's PD related to algebraic thinking. As they begin to sketch out a possible study, they note the need to identify content within algebra that is taught both in 7th grade mathematics and to 7th grade students taking algebra; they will also want to make sure that the 7th grade state assessment includes an algebra sub-scale that is well-aligned with the curricula used in the district. The team considers how to pick the teachers they will study, leaning towards a purposive rather than a random sample, selecting teachers who have similar profiles in terms of content knowledge but different patterns of student outcomes.

At the discussion continues, several team members suggest the possibility of a very different research approach, describing an idea they had talked about during the lunch break. Based on informal conversations with participants, they think that teacher beliefs about who can learn are the most likely explanation of the pilot year findings. They also think that teacher perceptions of principal support may be at play, as teachers may be feeling pressure to do drill and practice in order to raise student scores on district assessments. This group suggests modifying the intervention to address these issues, and then conducting research to see if that makes a difference, i.e., that teachers are willing and able to apply what they learn in the PD to their instruction in all of their classes.

These team members still suggest focusing on the 7th grade, for the reasons already articulated, but rather than studying a subset of pilot teachers, they want to do a broader study, including a larger number of 7th grade teachers. They suggest randomly assigning each of the remaining schools in the district to (1) receive the same intervention as the pilot group; (2) receive a modified version of the intervention with the teacher beliefs and principal support components added; or (3) serve as a comparison group, receiving the intervention a year later after the research has been completed. Observing a large number of teachers would be impractical, so the group suggests using a method for rating the implementation of instructional strategies based on a combination of teacher self-report and examination of selected classroom documents, and comparing the differences in the ratings of the two courses for each of the groups in the study.

The research team realizes they do not have the resources to conduct both studies, and decides to table the discussion for a week to give people time to consider the pros and cons of each approach, and then talk about it as a group again.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40.

Frechtling, J. with Mark, M. M., Rog, D. J., Thomas, V., Frierson, H., Hood, S., & Hughes, G. (2010). *The 2010 user-friendly handbook for program evaluation*. Rockville, MD: Westat. Retrieved from http://www.westat.com/pdf/projects/2010ufhb.pdf

Heck, D. J. (2008, March). *Applying standards of evidence to empirical research findings: Examples from research on deepening teachers' content knowledge and teachers' intellectual leadership in mathematics and science*. Paper presented at the Annual Meetings of the American Educational Research Association, New York, NY. Retrieved from http://www.mspkmd.net/papers/aera_march08_soe.pdf

Hill, H. C. & Shih, J. (2009). Examining the quality of statistical mathematics education research. *Journal for Research in Mathematics Education, 40*(3), 241–250.

MSP-KMD. (2010). *What we know about deepening teachers' content knowledge: Engaging teachers with challenging mathematics and science to deepen their content knowledge, Research on engaging teachers with challenging science content*. Retrieved from http://www.mspkmd.net/index.php?page=01_2d

National Research Council (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: The National Academies Press.

Sztajn, P. (2011). Standards for reporting mathematics professional development in research studies. *Journal for Research in Mathematics Education, 42*(3), 220–236.

Trochim, W. M. *The research methods knowledge base*, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/ (version current as of October 20, 2006).

Weiss, I. & Miller, B. (2009, January). *Determining what we know and how well we know it: The promises and perils of knowledge management*. Paper presented at the Math and Science Partnership Learning Network Conference, Washington, DC. Slide 65. Retrieved from http://www.mspkmd.net/papers/lnc_012709.pdf