Applying Standards of Evidence to Empirical Research Findings: Examples from Research on Deepening Teachers' Content Knowledge and Teachers' Intellectual Leadership in Mathematics and Science

For the AERA symposium: What Do We Know and How Well Do We Know It?: Methodology for Synthesizing Knowledge

Daniel J. Heck

Paper presented at the annual meetings of the American Educational Research Association, March 26, 2008, New York, New York

Prepared for the
Math and Science Partnership Knowledge Management and Dissemination Project
Horizon Research, Inc.
Education Development Center, Inc.

A Math and Science Partnership Research Evaluation and Technical Assistance project funded by the National Science Foundation

Acknowledgments

This report represents the efforts of several individuals and organizations. In addition to the authors of this report, many Knowledge Management and Dissemination staff at Horizon Research, Inc., Education Development Center, Inc., WestEd, and consultants from TERC were responsible for various tasks essential to the production of this paper.

This paper was prepared with support from the National Science Foundation grant number EHR-0445398. All findings, conclusions, and opinions expressed are those of the authors, and not necessarily of the National Science Foundation.

ABSTRACT

The Math and Science Partnership Knowledge Management and Dissemination (KMD) project deliberately developed standards of evidence that could be used for findings generated from qualitative, quantitative, or mixed methodologies, drawing heavily from previous efforts that had addressed only one of these categories. Although the KMD standards were developed prior to the release of the AERA Standards for Reporting on Empirical Social Science Research (2006), the underlying conceptual alignment is very strong.

Applying the standards of evidence is not intended to cast blanket "good/bad," "strong/weak," or "in/out" judgments on studies, nor is it to suggest that all studies should be strong on every standard. Rather the standards of evidence are meant to aid our understanding of the strengths and limitations of each study's contributions to the knowledge base for a particular topic of investigation. The standards of evidence were designed to facilitate a review process that examines the intended contributions of a study related to a particular topic, and to support judgments about those contributions based on the fit of the research design to the study's purpose; the quality with which that design was carried out; and the linking of findings to existing theoretical, empirical, or practical knowledge.

The paper describes the development of the standards of evidence, the process of their application to review empirical research studies, and the use of results from this review to summarize what is known, and how well, within a particular topic of investigation. The paper further provides examples from the review process and results from the project's work within the topics of deepening teacher content knowledge and teacher leadership. The paper concludes with recommendations for strengthening the knowledge base from evidence in empirical research.

INTRODUCTION

Math and Science Partnership (MSP) Knowledge Management and Dissemination (KMD) was funded as an MSP Research Evaluation and Technical Assistance project to support knowledge management within the Math and Science Partnership program and to disseminate information to the broader mathematics and science education community. The overall goal of the KMD project is to synthesize findings from MSP work and integrate them into the larger knowledge base for education reform. In this way, MSPs (both NSF-funded and Department of Education-funded) and the field at large can benefit from the research and development efforts of the NSF-funded MSPs.

The KMD project uses a three-stage knowledge management model, created by Nevis, DiBella, and Gould (1995) for workplace settings. The model posits that organizational use of knowledge occurs in three identifiable stages: *knowledge acquisition*, *knowledge sharing*, and *knowledge utilization*.

The KMD project is conducting its work in a few carefully selected areas of mathematics and science education research and practice that are central the MSP program and important to the field. Deepening teachers' content knowledge, and preparing and supporting teachers as intellectual leaders are two primary areas being investigated; the examples used in this paper are drawn from these two areas.

Reviewing the empirical research literature and situating the research taking place in MSP projects is a key task in the knowledge acquisition phase of the KMD project's work, resulting in summaries of research-based findings. Two other major components of KMD's knowledge acquisition include identifying key theoretical perspectives; and conducting interviews, panels, and document reviews to capture practice-based insights. These three sources—theoretical perspectives, research-based findings, and practice-based insights—are being integrated in syntheses that represent the field's knowledge base on selected topics and the contribution of the MSP program to the knowledge base.

This paper has three main purposes. First, it describes the rationale for the KMD project to develop standards of evidence to support its reviews of empirical research. Second, it outlines the process by which the KMD project applied these standards of evidence to review empirical research studies. Finally, some key results regarding the state of empirical evidence from two reviews are presented, concluding with recommendations for building on these bodies of studies to strengthen the empirical knowledge base.

Rationale for Developing Standards of Evidence

In the knowledge acquisition phase, the KMD project's charge for synthesizing existing empirical research was geared toward two basic questions: "What do we know?" and "How well do we know it?"

Answering the first question (What do we know?) involved an iterative process of developing logic models to situate research questions that might be addressed in studies relevant to a topic, searching research databases to identify studies that addressed the topic, summarizing the findings of each study, and pulling together findings that address the same research question or fit together within the logic model. Two primary screening criteria were used throughout this process. First, many documents were identified in the searches that were not reporting on empirical research studies. Empirical research studies were defined by systematic collection and analysis of data from a defined source or sources. Opinion pieces or advocacy pieces, which were common among the documents identified in the searches, did not fall within this definition and were consequently screened out. Second, to be included in the review, studies had to focus on a research question, either explicitly or implicitly, that informed the topic of interest. Studies that took place in the context of a program utilizing teacher leaders, for instance, but without a focused research question about this topic, were not included.

Once identified studies had passed these two screens, a set of common information was summarized for each study. The research questions or topics addressed in the study were delineated; the sources of data, instruments for collecting data, and methods for collecting and analyzing data were described; the empirically-based findings were listed; and the conclusions and implications drawn from those findings were identified. These individual study summaries, or, more particularly, the empirically-based findings or claims of these studies, were then grouped according to the specific relationships within the logic model that each study and finding/claim informed. These collections of findings/claims formed the foundation for answering questions about "what do we know?" within the topics of interest.

Answering the second question (How well do we know it?) required attention to the particular intent and quality of each empirical research study. More specifically, each finding in a study was assessed in terms the claim represented by the finding, and the evidentiary basis for making that claim. Education research has been under considerable and mounting scrutiny for the past two decades, leading to a number of calls for improved quality generally, and some proposed standards of evidence for education research to judge or improve quality (Century, Levy, & Minner, 2005; Howe & Eisenhart, 1990; NRC, 2002; Spencer, Richie, Lewis, & Dillon, 2003). A number of processes and tools for evaluating the quality of empirical research studies have been proposed, and many have been developed and are in use for various purposes. High profile efforts have emerged to define high quality education research, and to judge the empirical evidence resulting from studies, including the U.S. Department of Education's What Works Clearinghouse, and Great Britain's Evidence for Policy and Practice Information and Co-ordinating Centre. Education scholars have responded in support of, and with concerns about, these calls and efforts (e.g., Freeman, deMarrais, Preissle, Roulston, and St. Pierre, 2007; Kelly & Yin, 2007; Slavin, 2008), furthering attention to this issue and providing broader thinking and greater depth of understanding about what is considered to be high quality education research.

The KMD project's knowledge acquisition work has been directed toward synthesizing the findings from empirical research studies, in an effort to illuminate the "knowledge base" for specific topics. Additionally, the project has been concerned with the particular contributions of each study to the knowledge base. Describing these contributions led to careful consideration of the purpose of each study, the research frame and design that guided each study, the research

setting in which the study took place, the reported quality of the instruments and analysis that under girded the study, and the warrants for making claims in the form of findings, conclusions, and implications. A number of perspectives and tools for conducting this work informed the project's work, but none was designed specifically to do what the project was taking on, particularly for studies using a wide variety of methodologies, and potentially contributing to the knowledge base in quite different ways. For this reason, KMD developed a tool and process for applying standards of evidence to empirical research studies, with the specific intent to use this tool and process to assess studies according to their intended and realized contributions to the knowledge base.

Early in the project, a broad-based group of education researchers, methodologists, and education reform leaders met with the KMD leadership and staff to inform development of the standards of evidence for empirical research, including recommending existing resources to consult; members of this group later reviewed the tool and provided additional feedback. The KMD project's evaluators, researchers from the Consortium for Policy Research in Education at the University of Pennsylvania, provided additional critique. As the standards of evidence were initially being applied, the American Education Research Association released its *Standards for Reporting on Empirical Social Science Research in AERA Publications* (AERA, 2006), leading the KMD project to another internal review of the standards to ensure consistency with the AERA recommendations.

The resulting standards of evidence, and the process for applying them[1], result in a careful review of individual studies to provide ratings based on specific indicators, operationalized for different research methodologies, and narrative justifications for these ratings in six areas. The ratings allowed the KMD project to identify the strength of contributions of individual studies efficiently. For example, some studies provided thorough documentation of the programs they were investigating, contributing to the knowledge base for defining what it means to engage teachers with analysis of student work, or for a teacher leader to plan lessons collaboratively with other teachers. Other studies contributed to the strength of evidence that a particular program resulted in knowledge gains, or changes in practice. Still others might have been conducted in varying contexts to test generalizability, or have provided sufficient documentation of the context to characterize the limits of generalizability. The narrative paragraphs provided description and judgment regarding how the treatment of various indicator, as appropriate to the purpose of the study, influenced the strength of the study's contributions to the knowledge base and in what ways.

The standards of evidence, including the six major categories and specific indicators are outlined in Figure 1. The specific guidance provided to reviewers for three indicators is presented in Figure 2 to illustrate how the standards of evidence were developed for, and applied to, studies using different research methodologies, often because they were designed to make particular kinds of contributions to the knowledge base. Figure 3 illustrates the directions to reviewers for writing a narrative paragraph that provides evidence, appropriate to the purpose and methodology of the study, for one rating category.

[1] Dr. Daphne Minner of Education Development Center co-led, with the author, the development of the standards of evidence and the process for applying them.

Conducting the Review of Empirical Literature

KMD developed a system for conducting reviews of empirical research literature intended to ensure a transparent process with integrity and protections against bias in all phases. This process is outlined below in three parts: identifying studies for the review, summarizing and applying standards of evidence to the studies, and describing MSP-supported research.

Identifying Studies for the Review

The parameters for search and selection of studies for reviews are intended to yield a set of studies with a tight focus on the topic of interest. To be included in the review, each study had to meet all of the following criteria:

- The topic of interest (e.g., teachers' mathematics or science content knowledge, teachers' intellectual leadership in mathematics or science) was studied empirically, through a research question addressed using a specific measure or systematic analysis within the research design;
- The subjects or participants in the study were practicing in-service teachers within grades Pre-Kindergarten through 12, or in the case of teacher leaders, were practicing teachers in leadership roles who may have been released full-time; and
- The study was published since 1990.

A working team of researchers who have conducted studies on the topic of interest, along with KMD staff, identified seminal studies for the topic. These studies were located in the Education Resources Information Clearinghouse (ERIC), and the descriptor terms for each study were recorded. Search parameters for the review were identified by initially searching on these descriptor terms as keywords in ERIC to identify a larger beginning set of studies. Results of these searches were examined to identify additional studies that met the criteria for the review. Once identified, the ERIC descriptors for each of these studies were recorded. From these descriptors, the complete set of search parameters was developed and entered as a keyword search into two research databases. For the topics of deepening teacher content knowledge in mathematics or science, and teacher intellectual leadership, ERIC and the EBSCO Professional Development Collection were used as the two research databases.

For all documents identified in the final searches, a member of the MSP-KMD team read the abstract, and skimmed the study if needed, to determine its initial inclusion based on the criteria for the review. For both topics discussed here, close to 90 percent of the articles were eliminated in this initial screening. In most instances, the topic of interest was not studied empirically. Other documents dealt solely with pre-service teachers.

Summarizing Studies and Applying Standards of Evidence

Each included study included was summarized and reviewed by members of the KMD staff. The summary consisted of:

- An abstract, describing the study, including the methods or measures used to assess key constructs or variables;
- The findings of the study;
- The conclusions and implications of the study;
- A checklist describing the study's participants and the context in which the study was conducted; and
- The theoretical perspective(s) on teachers' mathematics/science content knowledge or teacher intellectual leadership represented in the study.

Summarizers also ensured that each study met the inclusion criteria based on their more complete reading of the documents.

Following the summary, each study was reviewed using a set of standards of evidence for empirical research. The KMD project developed the standards of evidence to operationalize principles for conducting empirical research in education and social science. The standards of evidence drew on numerous writings about research rigor, quality, and reporting, including efforts to address quantitative, qualitative, and mixed methodologies. A panel of mathematics and science education researchers, research methodologists, and mathematics and science education reform leaders assisted the KMD staff in the development of the standards of evidence to help ensure not only their quality, but also their broader utility.

The purpose of applying standards of evidence to the studies was to identify the contributions of each study to the field's knowledge base. Contributions were characterized in terms of what is known from the findings based on the intent and substance of the study, and the confidence that can be placed in the findings based on the nature and quality of the study. Applying the standards of evidence was not intended to make "good/bad" or "in/out" judgments on studies, nor to suggest that all studies should be strong on every standard. Rather the application of standards of evidence was conducted to aid understanding of the strengths and limitations of each study's contributions to the knowledge base.

Results of Applying Standards of Evidence to Empirical Research Findings

The efforts of the KMD project to address the questions, "What do we know?" and "How well do we know it?", from the empirical research on the topics of deepening teachers' content knowledge and teachers' intellectual leadership in mathematics and science have resulted in systematic reviews of numerous empirical studies within these two topics. The substantive results of these reviews, combined with the project's work on theoretical perspectives and practice-based insights, form the basis for more broadly based "knowledge reviews" currently being disseminated among the MSP community, and the broader field (See http://www.mspkmd.com).

The research reviews also provide a picture of the current state of the empirical evidentiary base for each topic—what questions have been investigated, in what ways, what claims are made, and what is the strength of evidence for these claims. Looking across the two reviews, and comparing the results to similar efforts to synthesize education research on other topics, points to some consistent and important patterns, from which recommendations for strengthening the evidentiary basis in education can be derived.

First, the findings in these studies are generally quite positive. Studies investigating interventions to deepen teachers' content knowledge or impacts of teacher leaders' activities on teachers' classroom practice, for instance, almost invariably find positive outcomes. The groups of studies in these reviews may be inherently biased, as publishing pressures (actual or perceived) may work against dissemination of studies that find either no effect or negative effects.

Second, the body of studies examining "downstream" effects of teachers' content knowledge and teachers' intellectual leadership (e.g., their relationship to classroom practice and student outcomes), supports claims that these topics matter. The empirical evidence in these studies, at minimum, points to these topics as areas worthy of attention in efforts to improve teacher quality and instructional practice. Despite limitations in individual studies and gaps in the collections of studies, there are fairly robust findings to indicate that, for teachers' mathematics or science content knowledge:

- Teachers' mathematics/science content knowledge influences how teachers engage students with the subject matter.
- Teachers' mathematics/science content knowledge influences how teachers evaluate and use instructional materials.
- Teachers' mathematics/science content knowledge is related to what their students learn.

Similarly, for teachers' intellectual leadership:

- Teacher leaders' practice, particularly in providing instructional support to teachers, impacts teachers' classroom practice.
- Teacher leaders' practice occurs in a larger context of conditions that impact teachers' practice.
- Teacher leaders' practice is related to student outcomes.

Third, studies that look "upstream" at interventions intended to deepen teachers' content knowledge, interventions intended to prepare teachers as intellectual leaders, or the effects of interventions implemented by teacher leaders, tend to be more like program evaluations than research. That is to say, they tend to examine the impacts of whole programs that may include a host of activities and opportunities. Many studies fail to identify potentially important aspects of the contexts of research, give little description of interventions and rarely investigate naturally occurring or planned variations in interventions to understand the relative importance and contributions of various features of professional development and leadership development programs. The typically positive findings of these studies offer existence proofs of programs that were effective, but generally leave unexamined which pieces of the programs, or combinations

of aspects, might be responsible for effects. Consequently, the program theories (Weiss, 1995) that posit linkages between particular activities or experiences and particular outcomes, although sometimes described, are rarely examined empirically. For example, a handful of studies of teacher leaders' work with classroom teachers included demonstration lessons or modeling, and each indicated positive results on teachers' classroom practice (Gersten & Kelly, 1992; Race, Ho & Bower, 2002; Vesilind & Jones, 1998). However, demonstration lessons or modeling of instruction was one of several practices in which teacher leaders engaged in these studies; the unique contribution or value-added of this strategy to effects on classroom teachers' practice cannot be teased out.

Fourth, many studies leave sample biases, response/participation rates, and attrition unidentified or unexamined. For instance, in looking at studies of interventions that engaged teachers with challenging mathematics content beyond the level at which they teach, participant groups involving solely teachers who had volunteered for an extensive intervention were common ((Basista & Mathews, 2002; Clark & Schorr, 2000, Garner-Gilchrist, 1993, Geer, 2001; Sowder, Phillip, Armstrong, & Schappelle, 1998; Swafford, Jones, & Thornton, 1997; Swafford, Jones, Thornton, Stump, & Miller, 1999), and some involved teachers who were screened for inclusion based on assumptions about the likelihood that they would benefit from the intervention (Garner-Gilchrist, 1993; Sowder et al., 1998). Comparison groups were not common in the studies, and some comparison groups that were included appeared to be initially different from the treatment groups. These limitations carry potential implications for both the validity/credibility and the generalizability of findings and conclusions. For example, studies intended to examine and document the potential impact of an intervention on teachers' content knowledge might appropriately involve providers who can be expected to implement the intervention with quality, and a participant group selected because they are thought likely to benefit from the experience. Some studies, however, were conducted under such conditions with an aim of assessing effectiveness, with little or no attention to limitations on the validity of claims that the intervention was responsible for the effects, or limitations of generalizability to teachers willing and able to commit to such extensive interventions.

Finally, documentation and measurement in the studies tended to be idiosyncratic. Documentation of programs to prepare teachers as intellectual leaders or to deepen teachers' content knowledge was highly varied, focusing on different aspects of the teacher leaders or teacher participants, describing different aspects of the interventions, and providing information on different features of the context of implementation. Also, few studies used existing instruments to measure teacher content knowledge or teacher leader practice. In many cases, especially when measures were developed specifically for the study, information on the reliability and validity of the instruments was incomplete or not reported.

Conclusions and Future Directions

Scholars in the field will continue to conduct many independent studies that contribute to the knowledge base in education. These studies might variously investigate interventions on different scales in terms of number of participants or substantive scope, evaluate programs or intervention strategies, compare naturalistic variations in treatment, describe the nature of

interventions or interpret participants' experiences from different perspectives, or something else. Regardless of purpose and methodology, it is a professional expectation in education research that scholars situate each study conceptually within the field and discuss its contributions to the knowledge base in which their work is situated. Certainly the knowledge base can be strengthened on a study-by-study basis through greater attention to this professional expectation.

Numerous reviews of research, including the knowledge acquisition efforts of the KMD project, indicate that a serious challenge remains for accumulating evidence across individual studies. Studies on deepening teachers' content knowledge and teachers as intellectual leaders in mathematics and science provide examples of programs, some well described, some not, that have shown evidence of effectiveness. At present, however, it is difficult to pull together evidence from these studies as they range widely in documentation and quality. It is particularly challenging to draw guidance from these bodies of evidence to inform practice in deepening teachers' content knowledge, or preparing and supporting the work of teachers as intellectual leaders.

Although some common features can be identified across these programs to formulate hypotheses about their relationship to measurable outcomes, the participating teachers or teacher leaders, the implementation of the programs, or the contextual circumstances may be unique. The instruments used to measure outcomes on which claims of effectiveness were based often have unknown properties or quality, and few were used in more than one study so that outcomes might be compared to understand what each program affords. Additionally, few studies were replicated to assess generalizability of results from a program across contexts, conditions of implementation, or participant groups. Burkhardt and Schoenfeld (2003, 9) highlighted some of the reasons that education research has tended to favor non-comparable studies:

Isolation of research efforts and lack of accumulation of evidence are evident broadly in education research (Burkhardt & Schoenfeld, 2003). These problems, in fact, are characteristic of efforts at organizational learning from social science research more generally, and are "exacerbated [by] lack of agreement on definitions and measures" (Huber, 1991, 9). In response, the National Research Council (NRC) has issued a broad statement about the current status and needed future direction of education research:

> Even if the quality of discrete education research projects has been ensured, if the field lacks the will or the tools to forge connections among studies, it will amass a multitude of studies that cannot support inferences about generalizability nor sustain the theory building that underlies scientific progress. We conclude that greater attention must be paid to reanalysis, replication, and testing the boundaries of theories with empirical inquiries, as well as to taking stock of what is known in areas of interest to education policy and practice on a regular basis. (NRC, 2005, 4)

The results of the KMD knowledge reviews have led to similar conclusions. Several directions for strengthening the evidentiary base might address these limitations. First, the field's efforts to improve the research designs, tools, and reporting in individual studies is a necessary step to

upgrade the quality of individual research studies. And, second, replicating promising programs under a variety of circumstances would provide a means to study the generalizability of results.

The goal recommended by the NRC of cumulating knowledge from collections of studies, or possibly from planned programs of research, suggests some additional directions for the future beyond strengthening and replicating individual studies. Based on the results of the KMD project's reviews of literature on teachers' content knowledge and teachers' intellectual leadership in mathematics and science, further suggestions to improve individual studies with an aim of fostering systematic accumulation of evidence are outlined below.

Individual studies can contribute to broader knowledge accumulation through their inclusion in research syntheses, including reviews and, in the case of quantitative studies, meta-analyses. Agencies and foundations funding research studies, or research centers supporting investigators with common interests, might be more or less directive in choosing specific topics and questions of interest. A more directive approach would fund or support independent studies around a common topic or question in order to generate sets of studies that are intended for inclusion in a research synthesis. A less directive approach would fund studies addressing a variety of topics or questions, leaving the work of combining results more open to chance. In either case, investigators can substantially expand the potential for knowledge accumulation through research syntheses by considerations in design, instrumentation, and documentation for each study they conduct.

Design Considerations

The most important design consideration for any study is choosing a design that is a good fit for the research questions that are being investigated (Howe & Eisenhart, 1990). For example, if the effects of an intervention, or variants of an intervention, are of interest, then the design should provide a means to ensure that effects are a result of the intervention or certain variants of it. The study design and implementation should either ensure that the intervention or the variants of the intervention are delivered as intended, or that deviations from intended delivery are well-documented. Studies should also provide a fair basis of comparison. Depending on the purposes of the study, comparisons may involve initially similar groups of participants, use of methods to account for initial differences, comparison to an explicit set of criteria or a shared experience of what is "typical" or status quo, or a critical examination of what is believed to be a common experience that may in fact be contextually or culturally dependent.

If the findings of a study are to be generalized appropriately, the design must provide the basis for that generalization. Characteristics of the research participants and how they were selected, features of the context(s) in which the research takes place, attributes of those delivering or supporting any interventions or other factors affecting participants' experience of interventions all affect generalizability of findings. For studies with generalizability as a goal, selection of samples, contexts, and providers/supporters that are representative of the conditions to which one wishes to generalize are imperative. In some cases, however, participants, contexts, and providers/supporters of interventions are not easily manipulated, or are determined by an existing environment or activity that is used as the opportunity for research. In these cases, documentation considerations are heightened, as described below.

Instrumentation Considerations

For independent studies to contribute to knowledge accumulation, the research instruments or methods for collecting data must be of acceptable and established quality. Measures of outcomes should be appropriate for use with the research participants, and valid for the purpose of drawing conclusions about the effects or relationships being investigated. Independent studies can establish appropriateness and validity for researcher-developed instruments, which in some cases are the only choice. However, by using existing instruments the findings of the study can more readily and more meaningfully be incorporated into research reviews or syntheses. Comparing or combining findings of studies that use common instruments is far easier than working from studies using different instruments that may not be measuring the same construct with similar quality. This point is not to imply that studies should not use a variety of instruments. In fact, comparing or combining findings that derive from a variety of instruments can be very valuable for examining the robustness of effects/relationships, or determining subtle distinctions in effects/relationships. Multiple measures used in the same study can afford these same benefits. The main point is that by using measures of established appropriateness and validity for the purposes of the study, the findings of the study are clearly enhanced and can contribute more meaningfully to research syntheses.

Documentation Considerations

Thorough documentation of studies, including samples, interventions, contexts, and methods, provides a basis both for interpreting findings from individual studies and for combining findings across studies. Even though many studies do not have designs that allow for systematic examination of the importance or effects of differences in background across groups of participants, contextual factors, or variations in an intervention, these factors can be examined in research syntheses and meta-analyses. The ability to do so depends on documentation in individual studies. Even when these factors do not vary within an individual study, documenting them affords greater contribution of the study to the field's knowledge base.

For example, in studies of interventions intended to deepen teachers' content knowledge, and interventions to develop teachers as intellectual leaders, thorough documentation would include information such as participants':

- Content, education, teaching, and leadership backgrounds;
- Teaching assignments, including grade levels and courses, and previous leadership responsibilities;
- Curriculum materials and other instructional materials and support resources in use in the district, schools, and classrooms;
- The policy content, including standards and assessments that guide or influence teachers' instructional decisions, and requirements or incentives for ongoing teacher professional development;
- Required or volunteer participation, and random or selective assignment to the intervention;

- School and district administrative support for the intervention and/or its goals for teacher learning and instructional change, or teacher leadership;
- The history of related improvement efforts in the school or district; and
- School and community demographics.

Documentation of interventions might include:

- Who delivered various parts of the intervention, including their background and preparation, and any other relationship they have 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 or teacher leaders will do as a result of participant, such as making instructional changes or providing services for other teachers; and
- Other professional development and/or leadership development participants have experienced, or other support they have received.

By documenting as much of the above information as possible, efforts to accumulate research through research syntheses and meta-analyses can aggregate, disaggregate, and analyze findings in a variety of different ways. The findings of individual studies can then contribute to knowledge accumulation, supporting broader learning about what types of interventions, and what strategies that make up those interventions, are effective in what ways, for whom, and in what contexts.

References

Basista, B. & Mathews, S. (2002). Integrated science and mathematics professional development programs. School Science and Mathematics, 102(7), 359–70.

Clark, K. K. & Schorr, R. Y. (2000). Teachers' evolving models of the underlying concepts of rational number. *Proceedings of the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, 22* .

Freeman, M., deMarrais, K., Preissle, J., Roulston, K., & St. Pierre, E. A. (2007). Standards of evidence in qualitative research: An incitement to discourse. *Educational Researcher, 36*(1), 25-32.

Garner-Gilchrist, C. (1993). Mathematics institute: An inservice program for training elementary school teachers. *Action in Teacher Education, 15* (3), 56–60.

Geer, C. H. (2001). Science and mathematics professional development at a liberal arts university: Effects on content knowledge, teacher confidence and strategies, and student achievement. *Proceedings of the 2001 Annual International Conferenceof the Association for the Education of Teachers in Science* .

Gersten, R., & Kelly, B. (1992). Coaching secondary special education teachers in implementation of an innovative videodisc mathematics curriculum. *Remedial and Special Education*, 13(4), 40-51.

Kelly, A. E., & Yin, R. K. (2007). Strengthening structured abstracts for education research: The need for claim-based structured abstracts. *Educational Researcher*, *36*(3), 133-8.

National Research Council. (2005). *Advancing scientific research in education.* L. Towne, L. L. Wise, & T. M. Winters, (Eds.) Washington, DC: National Academy Press.

Nevis, E.C., DiBella, A.J., & Gould, J.M. (1995). Understanding organizations as learning systems. *Sloan Management Review, 36*(2), 73–86.

Race, K.E.H., Ho, E., & Bower, L. (2002). *Documenting in-classroom support and coaching activities of a professional development program directed toward school-wide change: An integral part of an organization's evaluation efforts*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Sowder, J. T., Phillip, R. A., Armstrong, B. E., & Schappelle, B. P. (1998). *Middle-grade teachers' mathematical knowledge and its relationship to instruction*. Albany, NY: State University of New York Press.

Swafford, J. O.; Jones, G. A., & Thornton, C. A. (1997). Increased knowledge in geometry and instructional practice. *Journal for Research in Mathematics Education, 28*(4), 467-83.

Swafford, J. O., Jones, G. A., Thornton, C. A., Stump, S. L., & Miller, D. R. (1999). The impact on instructional practice of a teacher change model. *Journal of Research and Development in Education, 32*(2), 69–82.

Vesilind, E.M., & Jones, M.G. (1998). Gardens or graveyards: Science education reform and school culture. *Journal of Research in Science Teaching*, 35(7), 757-775.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), New approaches to evaluating community initiatives: Concepts, methods, and contexts. Washington, DC: The Aspen Institute.

Figure 1
MSP Knowledge Management and Dissemination Standards of Evidence

A. Adequate Documentation of Project Activities

1. Research question and constructs
2. Research site
3. Sample demographics
4. Interventions and implementation
5. Data collection

B. Internal Validity

1. Sample bias
2. Response bias
3. Attrition bias
4. Missing data bias
5. Contamination
6. Validity threats/Alternative explanations addressed through analysis
7. Validity threats/Alternative explanations discussed
8. Investigator bias/Reflexivity
9. Qualitative descriptive validity

C. Analytic Precision

1. Measurement validity/Logic of research process
2. Reliable measures/Trustworthy techniques
3. Appropriate and systematic analysis
4. Unit of analysis
5. Power
6. Effect size
7. Multiple instruments or sources of evidence
8. Multiple respondents
9. All results reported

D. Generalizability/External Validity Determination

1. Findings for whom
2. Generalizable to population or theory
3. Generalizable to additional contexts

E. Overall Fit: The extent to which the research questions, study design, data collection procedures, and analysis techniques align. Use information from Narratives A-D here for evidence to justify your rating.

F. Warrants for claims

1. Limitations presented
2. Decay and delay of effect
3. Efficacy
4. Conclusions/ implications logically drawn from findings
5. Conclusions/ implications grounded in theory

Figure 2
MSP Knowledge Management and Dissemination Standards of Evidence: Examples of Guidance for Selected Indicators

B. Internal Validity

8. Investigator Bias/Reflexivity	**Did the relationship between the *researcher*(s) and the treatments *bias* the results?** For example, did the researcher deliver the treatment or have a pre-existing relationship with the research participants? If so, were protections against possible biases employed? In qualitative studies, does the investigator have awareness "of how … interactions in a field site threaten, disrupt, create, or sustain patterns of social interaction [that] might result in a prejudicial account of social behavior in the site? Does the investigator protect against "individual preferences, predispositions, or predilections that prevent neutrality and objectivity"? (Schwandt, 2001). Some bias in terms of predispositions is inevitable, but it is important that the researcher engage in reflexivity, "the process of critical self-reflection on one's biases, theoretical predispositions, and so forth" (Schwandt, 2001).

C. Analytic Precision

2. Reliable measures/ Trustworthy techniques	**Were the data collection measures/techniques determined to be *reliable/trustworthy* for the groups/conditions under study?** For the purpose(s) of the study, were all the appropriate types of reliability (e.g. test-retest, internal consistency, alternate form, interrater, or agreement among independent coders) determined and results reported? If appropriate, were reported reliability coefficients at a professionally acceptable level, given the uses of the measures (e.g., lower reliability would be acceptable for examining group differences than would be the case for making decisions about individuals)? Were new instruments pilot tested? In qualitative data collection, were the trustworthiness and dependability of the data collection ensured, through strategies such as training for observation/interviewing, systematic adjudication of discrepancies among data replication of accounts by another researcher using transcription or video-taping, interrater checks on coding and classification, and clearly documenting the processes and procedures for recording and analyzing data.

D. Generalizability/External Validity Determination

2. Generalizable to population or theory	**Given the nature of the sample for the study, are the findings *generalizable* to a larger *population*?** Was the population from which the sample was drawn adequately described and was the sample chosen sufficiently large and appropriately selected to be representative of the population? For qualitative studies, is there evidence to support analytic generalization? That is, were cases selected in a purposeful manner to support, refute, or refine a theory (analytic generalization)?

Figure 3
MSP Knowledge Management and Dissemination Standards of Evidence:
Examples of Guidance for a Narrative Paragraph

Narrative paragraph for Adequate Documentation of Project Activities

This narrative should describe the extent to which there is a sufficient and clear description of the important indicators listed above, as relevant to the nature and purpose of the study. In other words, keep in mind whether the research question asked is exploratory, explanatory, etc. and whether the study is conducted in a quantitative, qualitative or mixed method manner.

The review must discuss whether sufficient information was or was not provided regarding:
–methods of data collection
–sample demographics
–constructs being operationalized in the study
–the intervention and its implementation (if an intervention was being studied)