WCER Working Paper No. 2009-4
June 2009

The Challenges of Producing Evidence-Based Claims: An Exploratory Study of NSF's Math and Science Partnership Community

Matthew T. Hora
Wisconsin Center for Education Research
University of Wisconsin–Madison
hora@wisc.edu

Susan B. Millar
Wisconsin Center for Education Research
University of Wisconsin–Madison
sbmillar@wisc.edu

Jessica Arrigoni
Wisconsin Center for Education Research
University of Wisconsin–Madison
jarrigoni@wisc.edu

Kerry Kretchmar
Wisconsin Center for Education Research
University of Wisconsin–Madison
kretchmar@wisc.edu

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

> Hora, M. T., Millar, S. B., Arrigoni, J., & Kretchmar, K. (2009). *The challenges of producing evidence-based claims: An exploratory study of NSF's Math and Science Partnership Community* (WCER Working Paper No. 2009-4). Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved [e.g., June 5, 2009,] from http://www.wcer.wisc.edu/publications/workingPapers/papers.php

The research reported in this paper was supported by a grant (DUE-0806280) from the National Science Foundation and by the Wisconsin Center for Education Research (WCER), School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency, WCER, or cooperating institutions.

Foreword

The members of the Math and Science Partnership (MSP) program staff at the National Science Foundation (NSF) are very pleased that this report, *The Challenges of Producing Evidence-Based Claims*, is now available for the MSP community and others interested in deriving evidence from K–12 intervention projects. An outgrowth of the MSP Learning Network Conference in January 2008, this report offers numerous findings and observations that remain salient as the original wave of MSP projects conclude their work and as new projects commence.

Since the MSP program's inception in fiscal year 2002, when we began with high expectations that a major federal investment would lead to important findings and models for the fields of science and mathematics education, NSF has stressed that projects need to contemplate and address what constitutes reliable "evidence" of outcomes for stakeholders, and this has inevitably resulted in considerations of project evaluation and education research. Our rich history with engaging the community on this issue includes:

- The inaugural Learning Network Conference in January 2003, just weeks after the first MSP projects were announced, titled "Building a Culture of Evidence";
- Requirements for each partnership project to complete a strategic plan, including an evaluation plan, within 180 days of the inauguration of its award;
- Funding of two Research, Evaluation and Technical Assistance (RETA) projects—Building Evaluation Capacity of STEM[1] Projects at Utah State University and Adding Value to the Mathematics and Science Partnerships Evaluations at the University of Wisconsin–Madison—to provide technical assistance in evaluation to the partnership projects;
- A workshop in October 2004 for principal investigators and evaluators of Cohort 1 and 2 partnership projects to formulate a statement that would guide effective project-level evaluation in the context of a national research and development effort (hub.mspnet.org/index.cfm/9924);
- Following the October 2004 workshop, discussions at the January 2005 Learning Network Conference and a considerable amount of additional work by a team of experienced evaluators, production of the document *Evidence: An Essential Tool* (www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf0531);
- Evaluation summits, framed by the two RETA projects, in September 2005 (hub.mspnet.org/index.cfm/11844) and October 2006 (hub.mspnet.org/index.cfm/13454).

Despite the above history, we—NSF, the MSP community, and the broader STEM education community—are still on the path of learning what constitutes evidence while working in the dynamic system of science and mathematics education. For the January 2008 Learning Network Conference (hub.mspnet.org/index.cfm/msp_conf_2008)—titled "Claims-Based

[1] Science, technology, engineering, and mathematics.

Outcomes: What Do We Know? How Do We Know What We Know? What Do We Still Need To Know?"—the call for abstracts required that submitters structure their abstracts by: (a) placing their work in the context of STEM education; (b) stating a claim or hypothesis being studied in some aspect of their MSP work; (c) outlining the design for the evaluation/research of that aspect of their work, including discussions of data collection and analysis; (d) discussing results from the evaluation/research and knowledge learned; and (e) offering conclusions and/or discussing implications of the findings. It was for this process and the subsequent conference that the MSP program engaged investigators from the University of Wisconsin–Madison to study how the MSP community responded to the call for abstracts and communicated evidence-based findings during presentations. *The Challenges of Producing Evidence-Based Claims* represents the fruits of that study and adds an additional component to the journey that we are on to strengthen our evaluation and research designs. Most recently, MSP program solicitations since fiscal year 2008 have increased the emphasis on education research in the MSP portfolio by offering original MSP awardees opportunities to conduct research through new Phase II awards and requiring that new Targeted and Institute Partnerships include a research component in their work.

Based on the findings, some of which influenced the direction of the January 2009 Learning Network Conference, there are two reflections that we, as members of the MSP program staff, would like interested readers to consider when they peruse the report. First, project teams must either use or develop methods that produce measures of evidence that are specific to their interventions and treatments, and both qualitative and quantitative methods are acceptable within the research and development environment of the MSP program. While studies that involve random assignment have gotten much attention in recent years, evidence in MSP investigations comes from multiple modes of analysis as projects transition their interventions from nascent intentions to more mature efforts. Second, student outcomes are not the only measure to consider. While eventual impact on students is of interest, project teams must also look for impacts that are proximal to their interventions and treatments. Moreover, sharing unintended and/or unsuccessful findings is useful and, indeed, important as part of the learning process.

Special thanks to Matthew Hora and Susan Millar for leading the study that resulted in this report and for stimulating those of us in the MSP community.

—Kathleen Bergin and Jim Hamos
National Science Foundation
Co-organizers of the 2008 MSP Learning Network Conference

The Challenges of Producing Evidence-Based Claims: An Exploratory Study of NSF's Math and Science Partnership Community

Matthew T. Hora, Susan B. Millar, Jessica Arrigoni, and Kerry Kretchmar[2]

Executive Summary

The goal of the National Science Foundation (NSF) Math and Science Partnership (MSP) program's January 2008 Learning Network Conference (LNC) on claims-based outcomes was to take stock of the work under way in the program and to explore the nature of evidence used to justify project claims. Aware that this conference could provide an excellent opportunity to gather data on grantees' evaluation and research accomplishments and challenges, NSF also commissioned a study, the findings of which are reported here.

This study describes the methodologies used by the MSP community to generate evidence-based claims[3] and seeks to understand topics of interest to the 320 LNC participants. The data set for this study included the 47 abstracts accepted for presentation, 68 interviews conducted during the conference, observations of all 26 breakout sessions, and 98 "think pieces" written by conference attendees. The analytic procedures included a holistic scoring rubric for the abstracts and inductive analyses of the interview, observation, and think-piece data using a structured approach to grounded theory. The study is framed by observed patterns in how principal investigators (PIs) and their teams responded to project evaluation requirements. Some PIs experienced a dilemma as to whether their dominant operational approach should be discovery—as for science, technology, engineering, and mathematics (STEM) research projects—or delivery of pre-specified outcomes. Other PIs and project leaders were slow to start the evaluation and were impressed by the complexity of producing sound evaluation findings.

We present our findings in three sections: (a) a synopsis of overall responses to the LNC, (b) findings from the abstracts submitted to the LNC, and (c) a presentation of primary factors influencing the MSP community's approach to evaluation.

[2] The authors—the Learning Network Conference (LNC) Analysis Group—wish to thank several people for their important contributions to this study. First, we thank the two National Science Foundation program officers, Jim Hamos and Kathleen Bergin, who engaged us to conduct this study and supported our efforts throughout the design, data-gathering, and analysis period. We especially thank Jim for his rapid, complete, and thoughtful responses to our requests and for his helpful suggestions and information. We thank the five members of the LNC Planning Committee who provided useful advice on the design of our data-gathering instruments: Nancy Bunt, Marilyn Decker, Thomas Dick, Ron Henry, and Marilyn Strutchens. We also thank the 10 members of Westat for their very professional contributions to the data gathering: Joy Frechtling, Amber Winkler, Joseph McInerney, Kimberley Raue, Gary Silverstein, Holly Bozeman, Amy Levitt, Gavin Fulmer, Molly Hershey-Arista, and Xiaodong Zhang. Special thanks go to Joy, who organized and led this superb group of Westat researchers and who provided insightful comments on a draft of this report.

[3] In this report, we use the term *evidence-based outcomes* to describe research findings or evaluation outcomes that are based on rigorous empirical studies of MSP project activities.

Overall Responses to LNC

Respondents generally indicated that they liked the conference *claims-based outcomes* theme and felt that it was timely and advanced the field in the right direction. Respondents were looking less for theory and ideal-world evaluation tips and more for realistic and field-tested ways to generate evidence and thereby improve their future efforts and build the field. Respondents also expressed diverse interpretations of terminology (e.g., *claims-based*, *evidence-based*), mixed reactions to the breakout sessions, and a persistent interest in the methods most appropriate for evaluating MSP projects. Finally, respondents conveyed a strong assumption that student learning outcomes were the outcome data of primary interest to NSF.

Findings from Abstracts Submitted to LNC

Our findings about the methodologies used by LNC participants to make evidence-based claims are based primarily on a rubric-based analysis of the 47 abstracts, and secondarily on interviews and think-piece data. The 2008 LNC was the first conference for which the MSP used the abstract format. The scoring rubric was developed after the abstract format had been provided to the projects. Overall, the findings were that the accepted abstracts had sufficient descriptors for each of the seven components of the call for abstracts. The mean holistic score for the abstracts was 2.8 out of a possible 4.0, with a range from 2.1 to 3.7.

Primary Factors Influencing MSP Community's Approach to Evaluation

We drew on all the types of data collected for this study to identify factors that support and challenge the MSP community's approach to evaluation. An understanding of these factors may enable the MSP community and others facing similar challenges to more effectively achieve the goal of producing rigorous evidence-based claims about the outcomes of education reform efforts.

We identified two supportive factors:

1. LNC participants, in general, clearly expressed the sentiment that evaluation is an important and useful activity for their MSP projects. The widespread interest in utilizing high-quality evidence explains in part the participants' interest in the LNC theme and conference proceedings. We view this interest—and the elements contributing to it—as a primary frame within which the following factors should be interpreted.

2. Several projects appeared to be using sophisticated evaluation designs and logic models that account for the nonlinear nature of change in educational organizations.

By contrast, we identified six factors that posed challenges:

1. Many respondents indicated that lack training in social science research, education research, and evaluation methods was a critical factor limiting MSP projects' ability to generate evidence-based claims.

2. Several respondents conveyed that their projects' evaluation activities were not adequately designed and implemented at the outset, forcing them to "figure it out as they went along" and compromising their projects' ability to make evidence-based claims about project efficacy.

3. Respondents encountered a number of difficulties in attempting to establish causal relations in complex research settings, including those associated with (a) evaluating multifaceted and evolving projects, (b) underestimating the challenge of collecting data from K–12 institutions, (c) working within the short time frame of MSP interventions, and (d) conceptually integrating the complex variables at work in their MSP projects in order to make evidence-based claims.

4. The effectiveness of implementer and evaluator interactions was uneven; while several respondents valued external evaluators and said that their projects were characterized by effective implementer/evaluator interactions, several others noted that interactions across these role groups were hampered by lack of planning, understanding, and effective communication.

5. While the MSP community was aware of national pressure to prioritize randomized controlled trials (RCTs) as the "gold standard" of evaluation methods, a few respondents stated that they believed that RCTs and other experimental designs were not feasible in many of the circumstances in which their MSP projects operated and expressed curiosity about other disciplines and methods that could be employed in MSP evaluation.

6. There was a widespread perception that the primary goal of evaluation was to show improvement in student outcomes. We speculate that various sources (whether real or perceived) have led many LNC attendees to hold this view, which has limited the range of evaluation methodologies and outcome measures that MSP project leaders draw upon.

Recommendations and Conclusions

We provide three sets of recommendations. The first includes steps the MSP program might take to improve future LNC conferences:

- Be cognizant of problematic terms like *claims-based outcomes*;
- Revise the LNC call for abstracts to specifically focus on research design;
- Focus breakout sessions more explicitly on evidence generation; and
- Consider using facilitators for each breakout session.

The second set pertains to strategies for improving project evaluation designs:

- Require proposals to demonstrate adequate evaluation planning and budgeting;
- Require PIs to demonstrate familiarity with basic elements of social science;

- Address limited notions of appropriate methodologies for evaluation;
- Emphasize the value of measuring intermediate effects, not just student achievement; and
- Require funded projects to establish internal procedures for evaluation data sharing between implementers and evaluators.

The final set of recommendations consists of strategies for improving the MSP community's expertise in evaluation methodology:

- Focus efforts on training and professional development workshops; and
- Explore methodological advances for conducting evaluation in complex institutions.

In conclusion, we congratulate the MSP program for commissioning this exploratory research project, as it constitutes an element of the program's own logic model for fostering greater rigor and transparency in education-related research and evaluation. We also emphasize that, although the study raises important questions and illuminates critical issues facing the MSP program, it was not designed to test specific hypotheses or to deeply explore specific topics. Finally, we suggest that the program conduct future studies that focus specifically on some of the factors identified here.

The Challenges of Producing Evidence-Based Claims: An Exploratory Study of NSF's Math and Science Partnership Community

Matthew T. Hora, Susan B. Millar, Jessica Arrigoni, and Kerry Kretchmar

I. INTRODUCTION

The NSF Math and Science Partnership (MSP) program aims to improve the teaching and learning of science, technology, engineering, and mathematics (STEM) disciplines while fostering mutually beneficial partnerships between K–12 districts and institutions of higher education (IHEs). Specifically, it encourages partnerships between K–12 districts and STEM and education faculty and administrators in IHEs in "efforts to effect deep, lasting improvement in K–12 mathematics and science education" (NSF, 2007). The MSPs are based on the premise that K–12/IHE partnerships should draw on the knowledge and experience of both K–12 and IHE personnel to develop strong mathematics and science content knowledge and pedagogical methods. The MSP program's theory of change is distinguished, in part, from prior NSF-funded programs by its expectation that the participation of STEM faculty in the teacher training continuum will result in lasting improvements in K–12 student learning (Change and Sustainability in Higher Education, 2006; NSF, 2007).

Since January 2003, the MSP program has been hosting an annual Learning Network Conference (LNC) for its grantees to provide a venue for the reporting and sharing of project findings and experiences in the field. The goal of the January 2008 LNC—titled "Claims-Based Outcomes: What Do We Know? How Do We Know What We Know? What Do We Still Need To Know?"—was to take stock of the work under way in the MSP program and to explore the nature of evidence used to justify project claims. This emphasis is consistent with the MSP program's focus on producing evidence-based claims in order to improve its programs and identify practices that could be replicated—and ultimately improve K–16 STEM education (NSF, 2007). It is also consistent with demands from the federal government and the education research community for greater rigor and transparency in education-related research and evaluation (No Child Left Behind Act, 2000; Kelly & Yin, 2007).

The MSP program was aware that the challenges confronting the evaluations of its 52 (now 73) projects had not been systematically identified and synthesized. With this situation in mind, and in line with its goals, the MSP program decided to obtain evidence on the extent to which its grantees were generating evidence-based claims. Accordingly, the program (a) designed its 2008 LNC to encourage the MSP community to understand its own progress toward making evidence-based claims about various types of project outcomes and (b) engaged us, the LNC Analysis Group, to study the 2008 LNC.

Study Design and Limitations

We designed this study to gather information about the methodologies used by LNC participants to make evidence-based claims and to systematically identify topics of interest to, and challenges experienced by, the MSP community. The data set for our study included (a) the

abstracts accepted for presentation (N = 47) that were provided before the meeting[4] and (b) data collected during the meeting, consisting of interviews (N = 68), observations of all breakout sessions (N =26), and think-piece exercises (N = 98) written by participants.[5] The data-gathering instruments and the LNC call for abstracts are included in Appendices A and C, respectively. A group of 10 researchers from Westat (see acknowledgements in footnote 2) provided invaluable assistance in gathering data during the meeting. The analytic procedures for our research included the use of a holistic scoring rubric for the LNC abstracts and inductive analyses of the interview, observation, and think-piece data using a structured approach to grounded theory (for details, see Appendix B). This report presents findings from this analysis, which we hope will be of value to NSF, MSP grantees, and the broader STEM education community.

We emphasize that our findings must be interpreted in light of four limitations. First, interviewees' self-reports, on which our findings largely depend, may be subject to social desirability bias; that is, interviewees' responses may be biased toward giving socially acceptable answers. Second, self-reported experiences recalled from memory may be filtered or modified by the very act of narration. Third, our sample represents a stratified sample of the entire MSP population, limited to those individuals who attended the 2008 LNC meeting. Thus, this self-selected population is not representative of the larger MSP community and includes disproportionate numbers of certain position types (e.g., project directors, junior staff, PIs, etc.). Finally, we used a semistructured interview protocol, which has both strengths and limitations. Its strengths lie in the validity of the data—interviewees freely choose to present what is most salient, important, or memorable to them, rather than responding, as in forced-choice questionnaires, to specific questions that a researcher believes are important. Weaknesses lie in the fact that respondents' *failure* to mention a topic does not imply that the topic was not relevant or of concern to them.

Because our thematic findings are based on comments volunteered by respondents, we cannot infer the perspectives of respondents who did not think to mention each specific theme. Thus, it is not appropriate to provide quantitative data at the level of detail that is justified when reporting on forced-choice responses. However, to provide readers an understanding of the *relative* numbers of respondents who volunteered specific points, we use the following verbal quantifiers:

2–3 respondents = *few*
4–9 respondents = *several*
10–30 respondents = *many*
All but a few respondents = *generally*

[4] A 48th accepted abstract, submitted by one of the authors of this report, was omitted from our analysis due to conflict of interest.

[5] The think-piece activity was designed to elicit descriptions of factors that affected participants' efforts to make evidence-based claims. Some 320 individuals (including presenters and observers) attended the 2008 LNC. Although all were asked, only 150 completed think pieces. Of these individuals, only 98 provided the kind of information needed for inclusion in this study. The 52 think pieces omitted from the data set provided only descriptive accounts of projects with no discussion of factors that affected project evaluation efforts.

For observations with only a few data points, we were especially careful not to generalize the finding to a larger group, and we included such observations only when respondents made particularly insightful points.

Note on Use of Quoted Material

In quoted material, we use brackets to convey in more grammatically correct or clear language what was stated by the quoted individual. Speakers are identified by their roles, such as *evaluator*, *researcher*, *PI*, or *project coordinator*.

Structure of Report

Below, we first provide a short summary of the larger context that motivates this study. We then turn to our findings, which include a synopsis of participants' overall responses to the LNC, an analysis of the abstracts submitted to the LNC, and an analysis of the primary factors influencing the MSP community's approach to generating evidence-based claims. We present our recommendations and concluding thoughts in the final section of the report.

II. THE CONTEXT: EVALUATION CHALLENGES

Since the mid-1980s, STEM and STEM education faculty—and a much smaller number of social science researchers—have acted as the PIs for education reform projects funded by NSF, other federal agencies, and private organizations. Most of these agencies, and in particular NSF, have asked these PIs to provide evaluation findings on their education reform projects. Over the years, independent evaluators who have worked with STEM education project PIs[6] have observed a pattern in how PIs (especially those trained in STEM disciplines) and their teams have responded to project evaluation requirements. We describe some of these patterns here.

One pattern is that STEM PIs often experience a dilemma of expectations. They are accustomed to undertaking research projects in domain sciences that pursue a discovery agenda, in line with NSF's clearly stated core mission. When engaged in discovery-based projects in their own STEM fields, researchers often state their hypotheses, describe the strategies and tools they will use to explore them, and speculate on the outcomes they hope to obtain. They also expect that the hypotheses and designs with which they begin may not hold. They know that many obstacles and opportunities they have not anticipated will be encountered as a result of their investigations. They therefore need—and avail themselves of—the freedom to adjust their design in real time, in hopes of identifying valuable new patterns or phenomena that can be verified and offered to the scientific community. However, these PIs often see education reform project proposals differently. Many such PIs believe that they must achieve the specific goals

[6] For example, one of the authors, Susan Millar, has been evaluating STEM education reform projects since 1990. She established the Learning through Evaluation, Adaptation, and Dissemination (LEAD) Center at the University of Wisconsin–Madison in 1994, and there helped design the evaluation components of some 175 STEM education reform proposals submitted to NSF and other foundations and agencies, of which 75 were funded and implemented. She developed several more STEM education reform grants after taking her position in 2003 at the Wisconsin Center for Education Research.

stated in these proposals. Education reform proposals present a set of envisioned goals that the PIs hope to achieve by using specified strategies that will lead to specific outcomes. And their evaluations are designed to track the success with which at least some of the specified strategies achieve the intended outcomes. Of course, as in any discovery process, STEM education reform projects encounter obstacles and opportunities that necessitate change in the planned strategies, which, in turn, may affect the feasibility of achieving the proposed outcomes. When this happens, PIs may experience the dilemma of needing to operate in a discovery mode, while also feeling—due to perceived or real pressures from different sources—that they are accountable for producing the outcomes specified in their proposal.

A second pattern emerges over the course of proposal writing and project implementation. During the proposal development period, PIs and other project leaders generally find the evaluation requirements challenging. This is not surprising, given that most are not trained in evaluation and relevant social science research and usually have not yet worked with evaluation professionals. Those whose proposals are recommended for funding generally have addressed this challenge quite well, either on their own or with assistance from their project director. Once funded, however, many find it challenging to produce credible and useful evaluation findings. Typically, their evaluation efforts are delayed, often because the project leaders (a) are preoccupied with early implementation efforts, (b) are not familiar with evaluation design processes (that typically require a theory of action and the development and collection of baseline measures), (c) are not certain which aspects of the project should or can be evaluated, and (d) are only weakly connected to an evaluator who could guide them. Moreover, many projects, especially in the 1990s, came to understand that they needed a larger budget for evaluation than originally anticipated. In many cases, PIs achieved clarity about what activities should be evaluated only after their projects had been under way for a year or more. Then, as they began to engage in serious efforts to carry out an evaluation, they often realized that it was a far more complex process than they had expected and looked for additional resources and assistance.

In an effort to help PIs address these common challenges and dilemmas—and more effectively undertake project evaluation—NSF has provided various evaluation resources (see, e.g., Katzenmeyer & Lawrenz, 2006; NSF, 2005). In addition, in time the MSP program came to require submission of evaluation plans for review by external evaluation experts, who provided critical friend feedback. The MSP program's decision to design its 2008 LNC to enable MSP leaders to present their best efforts to produce evidence-based claims represented one more step in NSF's ongoing effort to improve the capacity of the community. Another such step was their decision to commission the study on which we report here.

III. FINDINGS

Our findings are organized into three sections. The first draws on the data gathered during the LNC and provides a synopsis of participants' overall responses. The second draws on the abstracts submitted before the conference, complemented by interview and think-piece data, to describe how MSP projects generated evidence-based claims. The third and longest section draws on all types of data collected before and during the LNC to identify primary factors pertaining to project evaluation that members of the MSP community reported experiencing.

Overall Responses to LNC

This section includes a brief synopsis of participants' overall responses to the LNC. Respondents generally indicated that they wanted to see solid evaluations of their work so that they could continue to improve their projects and address NSF reporting requirements. Thus, they liked the conference theme and felt it timely. A primary factor behind this sentiment was the desire to discover new evaluation instruments and methods so that they could return home and pursue their work more effectively. Respondents also expressed diverse interpretations of terminology (*claims-based*, *evidence-based*) and concern about sessions that focused on implementation to the exclusion of evidence-generating methods, despite the emphasis on evidence in the call for abstracts. Finally, participants generally conveyed an assumption that student learning outcomes were the primary, and most useful, type of data.

Positive Responses

Approval for LNC Theme

Several respondents[7] had a positive reaction to the conference and its theme, *claims-based outcomes*. In general, people found the meeting useful and timely and obtained information from other participants. They expressed particular interest in practical ideas for generating evidence through evaluation activities and in hearing from people "in the field" and in "real settings." In other words, respondents were looking less for theory and ideal-type evaluation tips, and more for realistic, field-tested ways to generate evidence. A few respondents noted that this LNC was effective in achieving its goals and that they valued its focus on empirical evidence and its holding people accountable for producing solid data. A few respondents indicated that they had learned new methods and approaches. One respondent, a physicist, noted an aversion to making changes without data or evidence:

> [My interests] are what they have been forever as a physicist interested in teaching and learning—[namely,] without data, without information, and without analysis, you can't make improvements. I think that within our MSP we have made some initial steps toward basing what we're [doing] on real data. We haven't done any controlled clinical experiments yet. Those are hard to think of in education. Even harder than they are in medicine, but eventually we've got to get there. (Co-PI)

A few respondents explained that their interest in field-based efforts was motivated in part by challenges they experienced evaluating their MSP projects, including problems establishing attribution and causality in complex research settings (more on this below). These challenges were largely centered on establishing relationships between MSP interventions and student learning outcomes in response to perceived accountability pressures from NSF, the U.S. Department of Education, and Congress. For example, one participant conveyed this pressure by noting that because the MSPs are publicly funded, they need to show outcomes:

[7] Our sample represented a stratified sample of the entire MSP population composed of individuals who chose to attend the 2008 LNC meeting. We do not claim that this self-selected population is representative of the MSP community. See the Study Design and Limitations section (p. 2) for an explanation of the use of verbal quantifiers in this report.

> I think that they put that in our face a lot because of funding and because of political waves. And while I appreciate that it's not all we talk about here, it's always kind of hovering above us. (Administrator)

Perception That LNC Theme Provoked and Advanced the Field

A few respondents[8] noted that the conference theme was timely and welcome, especially as most MSPs are "wrestling with the same issues." In addition to appreciating the LNC's focus on developing the capacity to make evidence-based claims, a few respondents indicated that the theme provoked them to more deeply reconsider its implications. As one respondent put it, by highlighting the importance of evidence-based claims, NSF was attempting to "raise the bar and provoke those of us in projects to start thinking along these lines." A few indicated that they hadn't thought about evidence-based claims before the meeting or devoted significant time or resources to their evaluation activities.[9] Thus, the meeting forced them to consider these things.

> To be honest, this idea [of evidence-based claims] for us came out when the NSF posted a call for abstracts for this conference. We were not operating in that state of mind before, so we tried to accommodate our procedure in order to encompass this evidence-based abstract [idea]. (Project staff)

Another respondent observed that the emphasis on evidence and strong claims was forcing the partnership to be more focused in identifying outcomes and measures:

> It's had a profound effect in that we continually have had to reevaluate the way in which we look at outcomes, look at our assessments, and evaluate them at every step of the way. You know, consider what is a variable, what are the particular dependent and independent variables that we have, and determine where to put the validity measures and assign some level of validity to each of those—at least attempt to do so. So it had a lot of impact. (PI)

For a few respondents, the theme itself was not new, but it was evident that the MSP program was maturing in that over the course of the last three LNC meetings grantee methodologies were increasingly sophisticated. Indeed, one respondent expressed a desire for the community to develop even more sophisticated methodologies and approaches to well-known challenges with evaluation and research:

> I think we've talked about [these basic methods] before. I'm waiting to get beyond. The questions of what do we know, how we know it, and what else do we need to know are critical, and I think we have to keep asking them. But I think we're getting a little more sophisticated. I think there are times when we need to learn from doing and then the

[8] While we believe that the points reported here that were made by "a few" respondents provide potentially valuable insight into the experiences of members of the MSP community, these points should be interpreted with caution.

[9] We note that the groups of LNC participants who lacked familiarity with the MSP program's expectation that projects develop the capacity to produce evidence-based claims may have included individuals who had recently joined a project and who attended the LNC to obtain an intensive orientation to the MSP program.

> theory comes out of the doing rather than basing the doing initially on the theory. And I think we may have shifted too far the other way, at least in speech. (PI)

These findings indicate that members of the MSP community benefited from the learning opportunities provided by the LNC.

Appreciation for Community Building and Information Sharing

A few respondents brought up the role of community building in the LNC conference. These respondents found it valuable "just to see what other people are doing," to develop a sense of a professional community, and to share information. They viewed this community-building aspect as an important development, similar to the importance of bonds forged in a professional society through which individuals can share information and obtain advice from trusted colleagues. In particular, a few respondents felt relieved to know that their struggles were not unique.

> It's good to hear that we're not off by ourselves—the only ones having these kinds of issues. It's kind of reaffirming to see that other people are struggling with a lot of things as well. (PI)

In this regard, another participant noted that she valued being able to obtain the abstracts on MSPnet[10] so that she could learn what other people were doing.

Mixed Responses

Diverse Interpretations of Terms Claims-Based *and* Evidence-Based

Several respondents provided different interpretations of the terms *evidence-based* and *claims-based*.[11] One respondent, who felt that the terms were "sound-bites," expressed a desire for clarification from NSF:

> I'm not a scientist or an educator myself, and so I'm not sure exactly what that means and the fact that we've been through almost a full day and I'm still not sure what that means maybe means it hasn't been as much of a theme as you might think. (Project coordinator)

In particular, the meanings of *claims* and *data* were questioned, as these terms were perceived as being used interchangeably by the MSP program staff.

> At first it confused me because *claims*, to me, is different than *data*. *Claim* is just something you claim, so I think it would have been better to use *data-driven outcomes* [in

[10] MSPnet is an online resource for the MSP community that includes an extensive library of STEM education publications and resources, recent MSP evaluation reports, news and current events, and other resources.

[11] It is worth noting that the emphasis on the term *claim* derived from the Kelly and Yin (2007) article on structured abstracts, which in part inspired the theme of the LNC. Based on data collected for this study and extensive experience with STEM faculty, we conclude there is little evidence that STEM-trained faculty and staff are familiar with this use of the term.

> place of *claims-based outcomes*]. *Claims-based*—I guess that's just a friendlier term for data. (Director of research and evaluation)

Another respondent questioned the degree to which attendees' conceptions of the relationship between claims and evidence were shared. Observations of the breakout sessions supported that respondent's comment: observers noted that different interpretations of the terms were used. In particular, some presenters made claims based on their judgments about project accomplishments, rather than based on evidence derived from evaluation activities. A few respondents also noted lack of clarity about the meaning of these terms. For example, one said:

> I don't like the term *claims-based outcomes*. I've been asked by several people what that means. What you mean by that is very important, and I do wish more of the [breakout sessions] had been focused on evidence and how you collect and present that evidence. (Evaluator)

Another respondent conveyed that lack of clarity about these key terms was a serious matter because PIs pay attention to terms introduced by the funding agency. This individual indicated that his project's leaders had left a prior LNC with the belief that they were to collect data, even though they did not clearly understand how or why it would be used.

Mixed Reactions to Breakout Sessions

Research team observations and respondent accounts of the breakout sessions indicated that sessions varied in the degree to which (a) the sessions addressed the conference theme and (b) the audience became engaged. Regarding the first factor, the sessions generally split into two categories—those that were primarily evidence-based and those that were primarily implementation-based—as illustrated in the following observations:

> The presentations and audience participation were right on track with the conference theme. The nature of the breakout session strand allowed for great questions about procedures, questioning findings, and sharing ideas related to further MSP work. This was a very interactive and engaged audience, partly because the presenters truly facilitated the interaction and made their presentations engaging. (Research team observer)

> I do wish more of the conversations had been focused on the sort of evidence [you need] and how you collect and present that evidence, because sessions where that happens have been very interesting. That didn't happen all the time and indeed there were a lot of implementation presentations. (Evaluator)

Observers noted that 13 presentations explicitly addressed the claims-based conference theme in the presentation and/or the question-and-answer (Q&A) period, 4 focused to a limited degree on the theme, and 9 did not focus on the theme at all.[12] The presentations in the latter

[12] It is important to note that our observation protocol consisted of a general set of criteria developed by our entire analysis team. These criteria did not include a standardized process for assessing the degree to which individual presentations aligned with the conference theme.

sessions were largely descriptions of the project in its implementation phase, with little attention paid to evaluation design or evidence generation.

Consistent with the finding that the LNC participants generally expressed a strong interest in the LNC theme, a few respondents also expressed a desire to learn about how to collect and present evidence and wanted more emphasis on issues related to attribution and correlation. These attendees expressed disappointment with sessions that focused exclusively on implementation details or presented "data dumps" that did not explicitly discuss research or evaluation design issues.

The second factor that elicited mixed responses was the degree of audience engagement, which observers found varied significantly across the breakout sessions. Half (13) of the breakouts appeared to have sufficiently engaged their audiences, while the other 13 engaged them to a minimal extent. A few respondents also indicated *how* presenters successfully engaged the audience. For example, one noted that two sessions he had attended had the "right level of detail": they focused on context, but not minutiae, and thus were helpful to practitioners in the audience. A project administrator stated that he had taken away "something concretely useful" from two presentations and also was led to think about larger, more abstract issues facing his MSP.

Several respondents commented on the overall quality of the presentations they attended. One noted that overall the quality of the breakouts was superior to that at most professional meetings he attended, such as those of the American Educational Research Association (AERA). However, a few respondents expressed dissatisfaction with poor time management. For example, one respondent said:

> One of the things I believe is that people really did not leave enough question time. People presented for almost the entire period and left maybe 5 minutes at the end, and I think that's a real loss. I think that the intent was to have colleagues really talking about the research, and critiquing it, and helping each other, making suggestions, and I don't find that that happened. (Evaluator)

These respondents' comments are supported by the observers' reports, which indicated that presenters in the 13 sessions that did not focus on claims-based themes used all the allotted time, leaving little or none for discussion or Q&A. While this scenario is not unusual for presentations at professional meetings, respondents appeared to be especially sensitive to the issue at the LNC, as many were hoping to obtain practical insights into evaluation and evidence generation that they could then apply to their own MSP projects.

Methodological Considerations

Questions Relating to Effective Use of Multiple Methods

Several respondents raised issues about methodology that suggested divergent views in the MSP community on appropriate methods and procedures for MSP projects. For example, one perspective suggested a singular focus on obtaining student outcomes and "hard evidence," while another expressed an openness to and curiosity about methodological pluralism.

Respondents articulating the first perspective appeared to judge the adequacy of evaluation data in terms of their own academic discipline's approach to data. One STEM-trained respondent noted that he hadn't seen "hard evidence" at the LNC, only a lot of "interview data." The comments below also express STEM-based understanding of the nature of evidence:

> I think that it is imperative that this community produce research that follows a scientific model and that can be replicated, and I think it is very important to the credibility of this community and to the applicability of the results. (STEM consultant/project staff)

Respondents also commented on an apparent lack of receptiveness to learning about methods used in other disciplines. For example, a few noted the predominance of either quantitative or qualitative methods and wondered about the lack of pluralistic or mixed-method approaches to research or evaluation methodology. A mathematician who noted the preponderance of quantitative methodologies observed that the resulting data appeared unidimensional and wondered why MSP projects were not using more qualitative methods. Another respondent observed that other disciplines, such as economics, have potentially important methods and theories that could be used to enhance the methodologies used by MSP projects:

> Different ways of looking at data sets [are valuable], so for instance, our project is struggling, like a lot of other groups, to find out if [we've got] any kind of statistically significant results or if there's any statistical correlation at all. And there were a few coming from different areas, like economics, that provided us an insight and so [therefore we've] already started to make some connections with these other groups so that we can share our data and our methods. (PI)

A few respondents indicated that pluralistic approaches may help the MSP community deal with its significant methodological challenges, including the difficulty of assigning attribution and establishing causality in complex settings. One evaluator noted that a singular focus on student outcome data obtained using a single method was detrimental, explaining that a downside of using such an approach to evaluate complex projects like MSPs is that "there are some things that can actually be missed that are increments of improvement," and these increments may be important. Another evaluator noted the benefit of having multiple disciplinary perspectives:

> It was interesting, too, to see the folks who are STEM faculty versus the School of Ed folks versus the practitioners and to have those three perspectives in the room with this very new, not well-formed idea. How it was sort of picked apart from all those different perspectives was really interesting. (Project director)

One respondent noted the creativity in a session that highlighted the use of social network analysis as part of an evaluation:

> So it just has me thinking about whether it can be used in some way to measure cultural change or practice change at a school level or a district level. You know, like, some day might there be a social network analysis number that indicates a deep enough cultural

> change for lasting, ongoing collaboration or ongoing improvement. It's got more potential than anything else I've seen, to do that anyway. (PI)

Focus on Linking Interventions to Student Outcome Data

Many respondents indicated that they believed the primary goal of evaluating their MSP projects was to link project interventions to student outcome data, as opposed to other possible outcomes (e.g., improvement in teacher content knowledge, institutional change). As one respondent noted, "It's an issue we're all dealing with." In the view of several other respondents, the emphasis on student outcomes largely resulted from the mandate that the MSP program improve student outcomes and the strong focus at the national level on student achievement as the preferred measure of school effectiveness. For a few others, the conference theme pertained to finding tools to link project activities to student outcomes.

> That's what it's all about. It's "What are the students doing and how are they performing?" And if we don't address that aspect, we can do all of these cutesy little things, and we can engage in research projects—whatever we want—but if we're not really looking at the outcomes of student performance, then [what's the point]? (Administrator)

This focus on establishing causal relationships between MSP project activities and student outcomes, and the subsequent implications for generating evidence-based claims, is discussed in greater detail in the Factors Posing Challenges section and in the foreword.

Findings from Abstracts Submitted to LNC

This section presents information about the methodologies used by LNC participants to make evidence-based claims. This analysis is primarily based on the 47 LNC abstracts that the LNC Planning Committee accepted for presentation (see footnote 4), plus interviews, observation notes, and think pieces. Collectively, these sources provide a variety of data to describe how the MSP projects are addressing the challenge of developing evidence-based claims.

Methods

The LNC call for abstracts form (Appendix C) asked for abstracts that addressed the following criteria in no more than four pages: (a) context of the work to be presented, (b) claims or hypotheses examined in the work, (c) study design, data collection, and analysis, (d) results or knowledge claim, and (e) conclusions and implications.[13] Upon completing its review, the LNC

[13] It is worth noting that the call for abstracts, and thus our scoring rubric, were relatively vague in their criteria and did not include specifics about each category. As a result, neither represents as comprehensive a vision for evaluation designs as set forth in the two articles on structured abstracts (Kelly & Yin, 2007; Mosteller, Nave, & Miech, 2004) that were recommended in the original call for abstracts. These two articles include specific details about each criterion and also argue for designs that link research questions, data collection, and analysis in a coherent and logical whole.

Planning Committee identified several ways in which the abstracts accepted for presentation could be strengthened:

- Greater discussion of the scholarly context of the work;
- Enhanced discussion of the relationship of project claims to issues relevant to the MSP and/or the field;
- More detailed descriptions of the intervention, data collection methods, and data analysis procedures;
- More detailed description of the main findings and the data that support those findings; and
- Greater specification when drawing conclusions and making claims.

The LNC Planning Committee then requested that the authors of accepted abstracts revise their abstracts in light of these critiques, which 30 of the 47 elected to do. Based on the committee's feedback and the original call for abstracts, we developed a holistic scoring rubric to evaluate how the final abstracts complied with the planning committee's requests.[14] Our rubric was composed of the following elements: (a) general descriptions of the context of the work (including the specific research area related to the project), (b) the hypotheses or research questions, (c) the research or evaluation design, (d) data collection and analysis procedures, (e) the results or knowledge claim, and (f) conclusions and implications.

We analyzed each of the 30 revised and the 17 unrevised abstracts according to our seven criteria, using the following 4-point scale:

1 – Absence of descriptors (not included)
2 – Insufficient descriptors (included but lacking detail)
3 – Sufficient descriptors (included and detailed description)
4 – Exemplary descriptors (included and very detailed description)

We then derived a holistic score for each abstract representing the average of the seven criteria.

Overall Finding: Ambiguity About Adequacy of Designs

The mean holistic score for all of the accepted abstracts ($N = 47$) was 2.8, with a range from 2.1 to 3.7. These results suggest that, on average, the abstracts had sufficient descriptors for each of the seven components of the call for abstracts. In addition, in the interviews, a few respondents described their research and/or evaluation designs in ways that we consider methodologically sophisticated, rigorous, and commensurate with the criteria set forth in the MSP program solicitations. The following interview quote typifies these descriptions:

[14] We did not analyze the LNC abstracts in order to judge their quality. The planning committee selected abstracts for presentation at the LNC independently of the LNC Analysis Group.

> In [state name], we did a state-wide evaluation using experimental designs or quasi-experimental designs, which is [based on the literature]. Before we start implementing any programs, we pre-test teachers and their students, then we conduct the professional development program, and then we assess teachers and their students, all in both the program group and the comparison group. We measured the gain of teachers and students, and we're studying the results and we will publish them sometime within the next few weeks. (Administrator)

While this description is not generalizable to the entire MSP community, it indicates a methodological sophistication that was demonstrated by some interviewees and is substantiated by the abstract scores. That said, evidence from our breakout observations, interviews, and think pieces indicated that some MSPs were struggling with the challenge (described in the Context section, above) of developing and implementing good evaluation designs.

> Respondent 1 (Implementer): We got a late start, and we're just now gathering up the data. We're in our 4th year. We need another couple years to really do this right.
>
> Respondent 2 (Evaluator): And we've collected a lot of data, [but] we haven't sufficiently taken the time to analyze the data. I think there's a lot of data to be analyzed yet, and we didn't know how. I think we're getting a better idea of how now.
>
> Respondent 1: And we have changed the kind of data we're getting in some cases, too. So we've been struggling along, but we believe in the process.

Another challenge noted by a few respondents was that it is difficult to make claims that are tied to specific evidence in a rigorous manner. For example, one respondent noted that they had observed in the breakouts "lots of evidence, but it's not clear if it's the right evidence to make claims about success or failure for any individual project." Articulating a point that is well understood by evaluators—that in order to establish the relationship between evidence and claims, it is necessary to clearly determine a project's outcome measures prior to implementation—a K–12 administrator who is also an MSP co-PI made the following observation:

> We've learned that we have to start [designing our evaluation] before our project starts—that we need to find some way of measuring whatever it is you want to measure at the beginning, and that you don't wait. With the NSF project, we were maybe a year into our project before we realized that. (Co-PI)

These data indicate that some MSP projects were realizing that making claims that can be clearly linked to evidence collected is complex.

The ambiguity in our data about the adequacy of projects' evaluation designs prompted us to conduct additional analyses of the 47 accepted abstracts. We reanalyzed the abstracts looking specifically for references to evaluation design types (e.g., quasi-experimental, qualitative case study, or repeated cross-sectional), as these would indicate whether the authors had identified an established design, and hence a logic, for their work. This analysis revealed that 79% of the accepted abstracts did not name a specific evaluation design type or indicate how the various elements of the evaluation plan were causally linked. This finding had eluded us initially

because we had inferred the presence of evaluation designs from the generally detailed descriptions of data collection and analysis methods included in the abstracts. Most abstract authors described research design, data collection, and data analysis methods in a single section, as required by the LNC call for abstracts. In doing so, they generally focused on describing specific methods and analytic procedures without naming a specific type of evaluation design.

Since the call for abstracts did not explicitly require an explanation of how each element of an evaluation design was related to the others, an MSP could have strong data collection and analysis components yet lack or fail to present an overarching or strategic rationale, as illustrated by this interview quote:

> We had this myopic vision of the project because this was a professional development model [started in 2002] and not [an evaluation] model. So we were doing business as usual, doing professional development and not collecting the data [purposively]. Then the [NSF program officer] said, "You have a data collection model, but not an evaluation model." So it's taking us all this time to start thinking purposefully about the evidence that we're collecting. (PI)

The NSF program officer cited by this respondent was reacting to the project's exclusive focus on collecting participant data in the form of head counts with no attention paid to outcome measures. In the field of project evaluation, such counts are considered process measures that track the degree to which a project is being implemented according to its original goals. Ideally, these process measures are then linked in a logical fashion to outcome measures, in accordance with a theory of action that underlies the entire endeavor. This is often accomplished by using a logic model approach to evaluation design. In our view, this respondent collected process data but did not employ a logic model that linked process and outcome measures.

Specific Findings

This section presents detailed findings on the six specific elements of the LNC call for abstracts, based on our analysis of the 47 accepted abstracts. As explained above, we used a holistic rubric consisting of the following six criteria to identify how the abstracts complied with the planning committee's requests: (a) context (including specific research area), (b) hypotheses or research questions, (c) research design, (d) data collection and analysis procedures, (e) claims, and (f) conclusions. These criteria were based on the categories of the planning committee's call for abstracts. For the 30 revised abstracts, we were able to use the rubric on both the initial submission and the final, revised submission and document any changes that occurred.

Context of the Work

As previously noted, the LNC Planning Committee asked abstract authors to expand upon their original description of the context of the research to include more details about the scholarly tradition and/or literature base to which the project belonged. As a result, we scored each abstract using two categories for context:

- *Context I*: General descriptions of the background of the project

- *Context II*: Description of the specific research area in which the project operated

The mean scores for the abstracts ($N = 47$) on the Context I and II criteria were 2.9 and 2.6, respectively. The lower score on the Context II criterion is consistent with the LNC Planning Committee's observation that the abstracts generally suffered from a lack of detailed references to the literature base relevant to the project. By itself, the lower score for Context II does not provide a conclusive indicator of the degree to which MSP personnel were informed about the literature salient to their projects; it only indicates that they did not discuss the literature in their abstracts. However, the finding is consistent with others (discussed below) suggesting that many MSP personnel are working outside their home disciplines.

Hypotheses or Research Questions

The mean score on the hypotheses or research question criterion was 3.0. Examples of hypotheses and research questions include the following:

- Schools that exhibit strong distributed leadership characteristics will demonstrate stronger student achievement results in mathematics.

- What are the subject-specific needs of beginning math and science teachers at the middle and high school levels?

- Coursework designed to develop teachers' subject matter knowledge with accompanying professional learning community development will promote positive changes in secondary mathematics teachers' instructional practices.

It is important to note that, when assigning scores to this criterion, we assessed only the sufficiency of the description, not the *quality* of the hypotheses or research questions or their suitability to the particular research or evaluation design.

Research or Evaluation Designs

The mean score for the research or evaluation design criterion was 2.9. As noted above, when we initially scored this particular criterion, we incorrectly inferred that if descriptions of the data collection and analysis methods were present, then there was a research design. We did this in part because research or evaluation design, data collection, and analysis methods were bundled into one section in the call for abstracts. For our second analysis, we examined the abstracts for each of these three topics and found that 37 of the 47 abstracts (79%) provided descriptions of data collection and analysis methods but not design type. Of the 10 abstracts that did mention a specific design, 4 described a mixed-methods design, 1 described a posttest design, 3 described a case study design, and 2 used referenced descriptive or narrative designs.

Data Collection and Analysis Procedures

The mean score for the data collection and analysis procedures criterion was 2.9, which indicates sufficient descriptions of this topic under our rubric. As previously noted, the LNC

Planning Committee highlighted that the original abstracts lacked sufficient specificity on this topic. As the original score for this was 2.7, there was a slight improvement between the original and the revised submissions.

Results or Knowledge Claim

The mean score on the results or knowledge claim criterion was 2.9. While this result indicates a sufficient level of detail for this category under our rubric, the LNC Planning Committee highlighted this topic as problematic due to the use of generalities when making claims. As the original score for this category was 2.9, there was no change between the original and revised submissions. For both, we ascertained only the sufficiency of the description, not the *quality* of the knowledge claim or its degree of specificity. As a result, this analysis should be considered in light of other evidence. Other data sources indicated that some MSPs experienced challenges with establishing relationships between their data and results or knowledge claims, which corroborates the concerns expressed by the planning committee.

Conclusions and Implications

The mean score for the conclusions and implications criterion was 2.8. As previously noted, this topic was highlighted by the LNC Planning Committee as problematic due to the use of generalities when making claims. As the original score for this was 2.9, there was a slight decline between the original and revised abstracts. Again, as our scoring procedure did not consider the *quality* of the conclusions and their implications for the MSP, this analysis should be considered in light of other evidence.

Factors Influencing Evaluation Approaches

During our analysis, we drew on all the types of data collected before and during the LNC to identify the primary factors influencing the MSP community's approach to evaluation. Here, we present two positive factors and six factors that posed challenges, providing our own interpretations as we proceed. We believe that these factors are important to identify and understand, as they constitute the context in which the MSP program is unfolding on the ground and serve as supports or constraints for individual projects as they further develop their capacity to generate evidence-based claims. Because a few respondents noted that they were happy to see that others were "wrestling with the same issues," we believe that the factors presented here are influencing, and will be of substantial interest to, a large segment of the MSP community.

Positive Factors

Widespread Recognition of Value of Project Evaluation

LNC participants, in general, clearly expressed the sentiment that evaluation is an important and useful activity for their MSP projects. More than 90% of the participants who provided usable think pieces ($N = 98$) stated that their MSP's evaluation was an important resource for their projects. The primary reasons given for this sentiment included the desire to (a) obtain feedback on projects in order to make informed midcourse corrections (e.g., one project

manager wrote, "Findings from our MSP evaluation have been extremely useful in helping in make midcourse corrections to our implementation plan"); (b) challenge the assumptions and theories of action held by implementers; and (c) convey evidence of project efficacy to NSF.

In addition, some participants noted that evaluation could provide insights into the dynamics of STEM instruction and student learning at their implementation sites. For example, one participant highlighted the valuable role that evaluators can play as researchers as well as monitors of a particular project's efficacy:

> Particularly important to me have been the insights that our evaluators have shared on classroom practice. The district staff is very small, and the evaluators and researchers have given us a window into classroom practice that would otherwise not be available. (Co-PI)

One respondent, a staff person in charge of designing and facilitating K–12 teacher professional development workshops, noted that while it wasn't her "job to get the evidence to back the claim," she nonetheless wanted any evidence available indicating the ultimate efficacy of her activities. The widespread interest in utilizing high-quality evidence explains in part the participants' interest in the LNC theme and conference proceedings. We view this interest—and the reasons underlying it—as a primary frame within which the following factors should be interpreted.

Use of Logic Model–Based Designs and Validated Tools

A few respondents provided examples of individual MSPs that are meeting many of the challenges described in the Complex Organizational Settings section below by developing evaluation designs and logic models that specifically account for the nonlinear nature of change in educational organizations.

> What we did is begin with a logic model and we identified what the intended path to outcomes would be and we identified short-term, intermediate, and long-term outcomes. So even though we know student achievement is the main outcome, we want to be able to show important precursors so that at the end of the project, if we don't see [any changes] in student achievement, that we can at least say that it's plausible that at some point we'll see them because we see evidence of [change] in all these other things that are important to getting to [student achievement]. So for us, being able to articulate a logic model was absolutely key for being able to develop a causal thread of how we plan to get to outcomes using a certain set of professional development activities. (Evaluator)

We note that the use of logic models to operationalize a project's theory of action and to provide a framework for data collection and analytic procedures is consistent with good evaluation practice (W.K. Kellogg Foundation, 2004; Frechtling, 2007). In addition, we note that an evaluation intended to test for causal relationships among factors (e.g., professional development and student outcomes) should employ validated instruments from the literature in the appropriate field in order to save time and increase the validity of the findings. One PI illustrated another benefit of grounding a project in the relevant literature, explaining how

reliance on research-based instruments from fields in which the PI's project lacked expertise avoided the prospect of creating new and potentially invalid evaluation instruments.

> We made a decision to use the test for teacher knowledge instrument developed at the University of Michigan by Deborah Ball. I understand that that instrument is linked with gains in student achievement because of a study they did. (PI)

A comment by a respondent from another math-focused MSP makes evident that not everyone succeeded in finding relevant validated instruments:

> In our MSP, we are focused on mathematics and on increasing teachers' knowledge of mathematics for teaching. There are no tools out there for measuring what teachers have learned in terms of mathematics content. We did a big search in the first year and we've talked to people and looked at instruments, and they're not aligned with our project or with our PI's definition of knowledge of math for teaching, so that we've had to look for other ways to collect this data. (Evaluator)

Factors Posing Challenges

The MSP program believes that K–12 professionals and STEM and education faculty from IHEs must each bring their knowledge and skills to the table and learn from and collaborate with each other in order to accomplish the program's very challenging goals. Accordingly, the MSP program deliberately seeks to engage in each of its projects individuals from diverse disciplinary and professional backgrounds in order to create substantial new opportunities to learn and accomplish goals that otherwise would be out of reach. However, as the findings below indicate, working across disciplinary and professional communities poses challenges as well as opportunities.

Lack of Experience and Training in Education Research and Evaluation

Many respondents indicated that a critical factor influencing MSP projects' experiences with evaluation is that many project personnel lack training in social science research, education research, or evaluation. This lack of training is particularly problematic during the proposal development stage, as many PIs may have difficulty in determining which variables to measure, how to measure them, and how to design a robust research and/or evaluation project. Some respondents explicitly stated that they were experiencing some challenges because STEM faculty and K–12 implementers were not familiar with evaluation. For example, one co-PI noted that while he considered evidence-based claims "important," without any training in social science or education research it was very difficult to actually devise a robust evaluation plan that would generate reliable evidence. He explained: "Many of us are starting to do educational research without formal training." Familiarity with the basic principles of evaluation is particularly important for STEM faculty who play PI roles, as they are in charge of project design, implementation, and evaluation—activities that depend on types of methodological expertise based in the social sciences, not in the STEM disciplines.

As noted by a few respondents, the lack of training in relevant education research or evaluation means that project evaluation staff may not be aware of literature, instruments, theories, and history relevant to MSP activities.

> Most people in my MSP seem to be unaware or unfamiliar with research on student learning and with research-based instructional strategies. Educational research is therefore a second thought if it's a thought at all in the design of PD and other interventions and in evaluation. (Project director)

A few respondents also linked lack of training in education research with a tendency to adopt new ideas or innovations without the same skepticism staff bring to problems in their own disciplines.

> So I was in a group with the physics faculty, and [one of them] learned a new innovation in our pedagogy seminar and then tried it out. I would have been happy just to try it out, having learned about it from [pedagogy experts], and I would have felt that this would improve student learning. But she immediately thought to do a study on it and compile data and look at the data. It made me realize that I'm a little handicapped, having been trained as a mathematician in these issues. Now if we were to have a new grant I would have a much better idea from the get-go of how to develop a whole assessment data study. (PI)

Another respondent expressed that he was primarily an "advocate for kids," and the shift to "thinking as a researcher" was challenging:

> We understand the importance of [data]. We even are beginning to understand how to collect it. [As for] understanding how to interpret it and what inferences you can make once you have that data, it's very hard. And the people that I'm working with, none of us have had specific training. I did hire an outside consultant to work with us who has some background, done some research in data analysis and is helping us to learn how to think about it, but it's a much harder goal to achieve than people realize. (K–12 co-PI)

Finally, a few respondents observed that the MSP program's emphasis on evidence-based claims is providing professional evaluators and education researchers an opportunity to educate and train others who are engaged in STEM education reform.

> This culture of evidence is new to educators [on the project], and so they saw the evaluation design and research questions as external to their practice. Funders' insistence on more rigorous evaluation design and instrumentation has given evaluators a boost—and an invitation to educate [other staff] to show the value of [using] data to assess value. (Evaluator)

Inadequate Design and Implementation of Evaluation Plans

In light of the lack of training described above, it is not surprising that several participants noted that their evaluation activities were not designed and implemented as well as they would have liked. A few respondents indicated that their project's evaluation

implementation efforts were either slow to get started or inadequately designed, in some cases attributing these problems to the PI's lack of experience with education research and evaluation.

Some respondents acknowledged that they had not foreseen the time and money it would take to develop and implement an adequate evaluation plan and thus had not build it into their proposals. For example, one respondent noted that insufficient time had been budgeted for qualitative methods:

> We collected a lot of data but didn't budget time for analysis. Transcription and coding is time-consuming and expensive. (PI)

Another admitted a lack of expertise in data analysis:

> We have a lot of data, but we didn't know how to analyze it. We have been figuring it out as we go along. (PI)

Although it is not ideal for an MSP project to be "figuring it out as we go along," the situation is not uncommon, for the reasons stated above. Indeed, in situations where an emergent approach to project design and implementation is consistent with the discovery mission of NSF, some of these challenges may be unavoidable.

Complex Organizational Settings

Several respondents stated that evaluating MSPs was an incredibly difficult prospect because the venues in which MSPs operate (K–12 districts and IHEs) are "messy" and complicated. These respondents highlighted challenges in (a) establishing causal relations in complex research settings, (b) adjusting evaluation designs to evolving projects, (c) dealing with multifaceted interventions, (d) collecting data from K–12 institutions, (e) working within the short time frame of MSP interventions, and (f) making claims from collected data. While these challenges apply to all practitioners and evaluators, they are particularly salient for people working out of field, especially the STEM faculty who are generally the PIs of MSP projects.

We address each of these challenges below. It is also worth keeping in mind that underlying these challenges is a perceived imperative of establishing relationships between project activities and student achievement outcomes (see View of Student Outcomes as the Exclusive Goal of Evaluation section below).

Establishing causal relations in complex research settings. Several respondents highlighted the challenge of demonstrating cause and effect in complex educational systems.

> It's really hard to show cause and effect when you're dealing with a system that is so complex. We work with the whole system, all types of teachers and administrators, not just a few. The whole system doesn't change easily, and people saying they are making great changes have a hard time proving it. (Project manager)

Meeting this challenge requires comprehending all of the moving parts within the system, and also identifying confounding variables and controlling them (if possible or necessary). (We suggest that, in some cases, meeting this challenge may not be feasible.)

One by-product of this complexity is that some respondents are skeptical about the validity of their own projects' findings. For example, a STEM faculty member felt that the available data did not prove that the intervention led to the claimed results and stated that making evidence-based claims was practically impossible because "it's just messy." In addition, an evaluator noted that "there are too many confounding factors for us to really feel that anything is meaningful." One respondent identified the variability and movement within the independent variable as a key concern:

> As a scientist, I have difficulty with the variations that occur with the independent variable—namely, how to measure the amount and depth of [professional development] that a faculty member receives. We can see some changes in success rates of students, but relating these changes directly to interventions is difficult. (PI)

These observations underscore a core predicament facing the MSP community and those engaged in educational experimentation generally: how to establish causality between the intervention and student outcomes in a complex and multitiered organization.

Adjusting evaluation designs to evolving projects. A few participants noted another challenge with generating evidence-based claims for MSPs: MSP projects evolve and change in real time as the implementers figure out what precisely they are doing. In our view, such change is expected, as implementers are forced to respond to the unforeseen challenges and opportunities that will likely arise in the course of implementing projects in complex educational organizations (see NSF, 2005). MSP evaluations are methodologically complex and challenging, even in the best of circumstances. And the evolutionary nature of most MSP projects makes it even more difficult for evaluators to develop and carry out robust evaluation designs: planning for baseline measures and for data collection and analysis is much harder when the detailed strategies and tactics for achieving project goals are not clearly laid out in advance. One PI made this point by observing that trying to study and evaluate an MSP was like "building a bicycle while you're riding it . . . I mean, you're doing it and trying to figure it out at the same time." In our experience, such situations require that evaluators design their research in real time and engage in creative work-around strategies. As noted previously, this research and development aspect of the MSP program presents unique challenges.

Dealing with multifaceted interventions. Another challenge noted by several respondents is dealing with the multifaceted nature of projects as they unfold in complex organizations. One respondent explained why it is difficult to generate evidence-based claims under these circumstances:

> We have a project that is multitiered, including instructional materials, math leaders in every building, professional development for 575 teachers, and training principals. And so we have all of these interventions, but we can't point to any single intervention [as the cause of changes in instructional practice]. There are so many factors involved with what we're doing that it's really hard to make any claims at all. (Project manager)

In our experience, the challenges posed by such multifaceted interventions surface as the *unit-of-analysis problem*, which has bedeviled organizational researchers for decades. In targeting different points in a complex system (e.g., individual teachers, departments, and entire organizations), implementers and evaluators must decide if and how they can make claims about changes at multiple levels based on their available evidence. The PI quoted below makes this point, while also remarking on the value of the LNC conference:

> I have been very interested in thinking about claim-based outcomes at different levels. So when the unit of analysis is the teacher, I find that I can have discussions within our group about that. Whenever we try to get to the school, district, and the program levels, I'm finding that I'm not feeling like the causality is strong enough [at these levels] for one particular intervention. I do think that these conferences allow you to think more deeply about that and how your impact fits into the bigger whole. So to think about how other people are dealing with those kinds of questions is useful. (PI)

As this observation indicates, some MSP evaluations are not designed to collect evidence that would warrant claims of project effects at multiple levels. Nonetheless, some respondents have "messy" evidence of such effects.

> What we do know is when you put all of these pieces together, it's a rich [collection] of different kinds of interventions, and there seems to be improved instruction because we can point to [classrooms] that had made massive improvements. [But] it is just messy. (Project manager)

In our view, the challenge of working at multiple units of analysis underscores the importance of understanding that an effective evaluation design does not attempt to accomplish more than is feasible, given the nature of the project, its timeline, and its resources. That is, many projects encounter difficulties because they underestimate the challenges of accounting for the complex and multilevel nature of educational organizations.

Collecting data from K–12 institutions. Several grantees explained that they faced challenges because partnering K–12 districts had proven reluctant to provide student- and teacher-level data. A few respondents attributed this reluctance to the "large and bureaucratic" nature of the districts, which have difficulties simply collecting data, much less disseminating it to other organizations. Right-to-privacy policies also affect access to data.

Another challenge is that sometimes state mathematics and science testing policies and procedures change in the midst of an MSP project, which complicates the task of comparing test scores collected over time.

> The biggest factor in implementing the evaluation has been changes in state testing in math and science, both in the content and changes in the grades given the test. In [our state], student testing changed midstream. (PI)

Unanticipated difficulties collecting teacher-level data are associated not only with districts' refusal to provide the data, but also with teachers' changing grades, teaching more than one subject in a given year and over time, and participating in multiple professional development activities at a given time. In addition, a few participants felt stymied upon realizing that they had

developed their evaluation designs based on what turned out to be erroneous assumptions that certain types of data would be available. This point was made by the following think-piece writer:

> In many cases, we have not been able to collect the data with the instruments developed, and much of the existing data is accessible in a manner that does not support the intent of the project. Many changes and deletions have had to take place within the evaluation plan. (Evaluator)

NSF requires projects to submit baseline data as part of the proposal submission process. However, to ensure the availability of needed data, MSP projects would be wise to go further once funded, determining the scope of available data in their partner K–12 districts prior to designing the project evaluation. In addition, a strong practice during the proposal preparation stage would be for partners to be explicit about data desires, needs, and expectations.

Working within a short time frame. Several participants ventured that the time frame of most funded projects may too short to demonstrate project effects on student achievement. The following statement is representative:

> It is very difficult to tie the professional development to student outcomes. It takes time to work with teachers to upgrade their content knowledge and skills, yet we are expected to show changes in outcomes almost immediately. (State coordinator)

As several MSP projects are using teacher professional development as a strategy to effect changes in student achievement, the potential lag time between professional development, subsequent instructional changes, and effects on student achievement is a significant issue.

> [It is] a big challenge—a lot of these projects have to do with teacher professional development and there is a lag effect. Do you have professional development and *bam!* the teacher goes into the classroom and is changed, or does that develop over time? (Project manager)

Despite this challenge, a couple of participants expressed some optimism that in the long run their MSP projects would be able to establish causal relations between professional development and student outcomes.

> It's too big a leap to link professional development and student outcome data. One or two years into the project is pretty early to look at student outcomes. Five years might actually say something. (Administrator)

> I think at this point certainly, the passage of time in terms of being able to look at things longitudinally is promising. People are really grappling with [these] questions, and I think now projects have multiple years of data, and the scope of the work that people are able to do is enhanced at this point. I think a lot of that is the projects are more mature and people have learned sort of along the way about how to look at these things. (Project manager)

The MSP program's recent introduction of Phase II opportunities for initial awardees will enable some MSP projects to conduct focused research on some of their interventions 6–8 years after the projects began.

Making claims from collected data. Several respondents focused on difficulties they had in using the evaluation data they had collected to make credible claims—a difficulty that reflects the challenges of working in complex organizations and using evaluation designs that do not specify ahead of time the logical relationships among the research questions, data collection, and data analysis procedures. One PI described this problem as follows:

> There's a lot of evidence, but it's not clear exactly that it's the right evidence to make claims about success or failure for any individual program. There's just way too much information and far too little evidence. I think there are lots and lots of numbers that are collected, there are lots and lots of surveys that are given out, and there is important information, but it all needs context in order to understand exactly what it's trying to tell you. (PI)

In addition, when the frame of reference is the "basic sciences," respondents may have particularly high standards when it comes to making evidence-based claims, as illustrated in this respondent's comment:

> For someone who comes from basic sciences, *evidence* is a very strong word. We don't have an awful lot of evidence in terms of what we're really doing here. There's an awful lot of information, and some of it is useful—quite a bit of it, actually—but I wouldn't claim that it's scientific evidence for very many conclusions that we might want to draw. (PI)

We found that this type of skepticism about claims was expressed by respondents with education identities as well.

> I find it really hard in this kind of work to really say, "Well, because we did this, we got these results." Like that connection I'm finding, the data that I saw doesn't necessarily prove that. There definitely are some improvements [needed], but how do you know? And I don't know that there's a good answer to that. I just think it's messy. (K–12 personnel)

This skepticism should also be considered in light of the previously noted tendency for some LNC presenters to make claims about their success by describing their implementation efforts, rather than by presenting data that substantiates outcomes.

Finally, some MSP projects assessed the complex nature of educational organizations in which they were operating and chose to focus instead on smaller, more discrete aspects of a either a school or an IHE.

> Well, I could say that it's really hard to get claims-based evidence. If you ask important questions, it takes a long time and a lot of different data to collect. So we've seen one small piece of evidence in this breakout session—social networking [which tells you

> about professional community]. It's hard to get all the pieces and then put it together in a big picture. (Evaluator)

This respondent reinforced the notion that conducting evaluation in complex organizations is extremely difficult, both conceptually and methodologically, and noted that it may only be possible to get credible evidence about limited elements of the system.

As a whole, the above comments demonstrate the tension that occurs when natural scientists steeped in experimental design tackle social science, in which qualitative or correlational studies with statistical controls are conducted. These respondents appeared doubtful that such studies could result in valid information about what was being learned. Furthermore, many respondents expressed skepticism that conducting research in naturalistic settings, where many factors cannot be controlled, could yield strong evidence of the sort obtained in the laboratory-based contexts with which they may be more familiar.

Implementer-Evaluator Dynamics

In line with the desire to utilize evaluation data for project improvement, many respondents discussed the nature of the interactions between MSP project implementers and evaluators. These observations generally focused on implementer and evaluator interactions and their implications for generating evidence-based claims.

The value of external evaluators. Several respondents highlighted the value of hiring objective evaluators who are external to the project, a practice mandated by the MSP program. For example, one respondent raised the issue of implementers' lack of expertise with evaluation and research design as a primary rationale for hiring external evaluators:

> I don't think [the project implementers] knew the expectations [for evaluation], so they may know how to define what they want to do and the outcomes they want, but they don't necessarily know how to evaluate [their project]. [And] I don't think that they're trained to do that, which is why external evaluations are really important. (Administrator)

Another respondent described working with external evaluators as "far more beneficial" than working with internal evaluators. This respondent then described the benefits of working with external evaluators:

> They were responsible for all [evaluation] activities, allowed us to examine more phases of the project, reduced bias, and provided formative evaluation at key points in the project. (Project director)

We learned that a few MSPs worked with external evaluators to collaboratively develop evaluation designs and instruments at the beginning of the project. In one case, the PIs did not want "generic observations or protocols," so they worked closely with their evaluators to identify evaluation instruments that were tailored to their particular MSP. This collaboration was inspired, in part, by a desire to obtain usable evaluation data that would inform later project activities.

> We worked with the [external evaluators]. They came up with the design, but we had input on the instruments. We didn't want generic observations or protocols because we wanted to get guidance for future work out of the results. (Project director)

In another case, an unsatisfactory NSF site review led to a renewed focus on evaluation, and close collaboration with external evaluators resulted in development of project-specific instruments.

> After the Year 2 site review, where we were really kind of [dinged], we [realized that we] had to have a research piece to MSP. We did a very good job of identifying a set of studies that external evaluators would conduct to not only identify the results of our work, but also to provide guidance in any kind of future work that we're doing. So we identified the studies ourselves and then we met with each of the [three] external evaluators. They came up with a design, but we gave input to make sure that the survey instruments, the observation instrument, [and the] interview instrument were directly related to the work that we were doing in MSP and not some kind of generic observation or survey instruments. They have to be tied with what we're doing. So we gave a lot of input at the beginning in order for this to occur. (Project director)

In our view, given the widespread desire by several respondents to use evaluation findings, this type of collaborative design process may be of value and interest to the broader MSP community.

In contrast to the pattern described above, some PIs completely delegated responsibility for designing and implementing the evaluation to external agents. In one of these situations, a PI later regretted that decision:

> In hindsight, there are places or ways that we could have more carefully collected data. We were caught up in doing the work of the project, and we left the evaluation to [the evaluators], which they did a great job at—a job that we could not have done. But maybe in future projects we might set aside more [implementer time] to do it. (PI)

Another view expressed by a few respondents was that some implementers perceived evaluation activities as "external to practice." In these cases, it appears there was a complete separation of the two activities, such that project implementers and evaluators did not effectively communicate or coordinate their respective efforts.

While evaluators need to be removed from the implementation of a project in order to maintain the integrity of the evaluation process, we believe there should be an active feedback loop between the two parties. When this is not the case, implementers are much less likely to use the evaluation data. Furthermore, given the challenges with obtaining evaluation data, as discussed above, it is doubly important for evaluators and implementers to communicate effectively in case changes in evaluation design and/or protocols prove necessary.

Communication issues. In speaking about interactions between MSP grantees and evaluators, several respondents underscored the importance of communication. One issue raised by a few respondents was the significant lag time between evaluators' collection of data and their

presentation of findings to the implementers. Since these respondents viewed evaluation data as an important resource for their MSPs, they thought this delay was unfortunate.

> We invest a large amount of time and money in our evaluation process, and the data is collected over a long period of time, [but] there's this big lag between when you've collected it, when you get the report back, and what that actually means as you are continually moving forward and implementing cohort after cohort. I'm not sure at this point that what we've gotten is useful to informing our own project. (PI)

From the evaluator perspective, communication issues often emerge during the proposal writing process. One respondent described the importance of this initial stage of collaboration and the implications that a poorly designed evaluation may have on the project's ability to generate evidence-based claims:

> As the evaluator in this project, I have been working on convincing the project staff that the goals and objectives of the project are more than just guidelines for evaluation. When we put in the proposal, I cautioned that if the project had student achievement as a goal, then the project staff would have to [design] the evaluation accordingly. This has been difficult [in practice], as project staff have insisted on project-specific assessments. Promising a quasi-experimental design means working to deliver one, and project staff has struggled with this interpretation. A goal/objective-focused evaluation has been a struggle with the mindset that project evaluation means just describing what the project did—a previous mindset of the co-PIs. (Evaluator)

In our view, this observation speaks to the larger issue of the different areas of expertise held by MSP implementers and evaluators, a matter that is particularly important during the design phase of a project.

Methodological Pluralism

As background to this next factor, we note the presence of a trend in education research to consider randomized controlled trials (RCTs) the "gold standard" of evaluation methods (U.S. Department of Education, 2003). While the MSP program has not encouraged or discouraged RCTs, this trend can be considered a backdrop to the MSP program. More recently, the Academic Competitiveness Council (U.S. Department of Education, 2007) developed a hierarchy of study designs that placed RCTs at the top, asserting that they most effectively establish impact on student learning and ensure accountability for investments (U.S. Department of Education, 2007). The council's decision was based, in part, on the finding that few educational programs are rigorously evaluated and that impact on students is poorly understood (U.S. Department of Education, 2007).

A few of our respondents expressed the view that RCTs and other experimental designs—while exceptionally well suited for establishing causality between an intervention and a dependent variable—are only viable or suitable for particular research settings.

> *Claims-based evidence*—when I hear that kind of term, it strikes me as leaning along the lines of what the DoE would call the gold standard of evaluation, with randomized

> control trials. And I don't think that that is necessary to show impact. I know that this is maybe simplistic, but if it looks like a duck and walks like a duck and sounds like a duck, it's probably a duck. So there's a lot from the evaluation side that face validity has a lot to do with things. I think that there's a lot of face validity being demonstrated here, but the skeptics will not [accept it]. (Evaluator)

As we interpret it, this respondent did not accept the idea that RCTs are the only way to generate credible evidence and implied that methodological pluralism is needed. We note that this respondent's sentiment is corroborated by the American Evaluation Association (2003) and some professional evaluators (Patton, 2006).

Other LNC attendees also voiced this view. For example, in the discussion period following one of the keynote addresses, participants commented that conducting an RCT in a K–12 school district is very expensive, labor-intensive, and wrought with methodological challenges. Another respondent pointed out that, while RCTs might be feasible in some STEM fields, there are diverse reasons why they may not be feasible in many of the circumstances in which education reform is under way:

> Well, some things that are actually common in science are, for any number of reasons—sociopolitical reasons and just philosophical reasons—not doable for the projects that we're in. Randomized blind tests, for example, we simply can't do. (PI)

Others explained that education is a field that poses unique challenges—including participant attrition and multiple confounds—to the use of RCTs, as well as other methods. For example, another respondent suggested that instead of aiming to demonstrate a causal relation between a specific intervention and student outcomes, other alternative methods, constructs, and measures should be considered:

> Repeatability is very hard to achieve from year to year, [meaning] you have someone who's doing pretty much exactly the same thing they did the year before, but you've changed one variable. That's very hard. We tend to change several things at once. You're trying to improve programs. And I think those [challenges] are almost insurmountable. So you might direct the evidence-based [research] towards things that are relatively easy to [prove], where there are common agreements on these things and where we've done enough study on it and there's enough understanding that you can get to a common belief that there's no real difference if you change this or that. [Maybe] the [interventions] are based on learning theory, psychology in the classroom, whatever. You just sort of want to take as the underlying basis a common belief [in the field] and then you modify those things across programs and then see what the effects are. Then, if you see something that's relatively common and it's statistically significant, I'd say you have a reason to say that there's evidence. (PI)

View of Student Outcomes as Exclusive Goal of Evaluation

Our data indicate that several respondents perceived that the primary goal of the MSP program—and the rationale for their evaluation efforts—were to provide evidence of improvement in student outcomes, particularly test scores. As previously noted, sources of this

perception include the No Child Left Behind Act of 2001 (NCLB; 2002), the push for accountability in education, and pressure to achieve greater methodological rigor in education research and evaluation. We speculate that various sources (whether real or perceived) led many LNC attendees to state that their evaluation efforts were focused on student outcomes.

> And I think that's the discussion: How are you going to judge the effectiveness of a project? That's where I'm interpreting that the effectiveness of the project has to be [judged] on student performance. (Administrator)

Many respondents repeatedly commented on the challenges related to their lack of knowledge about education research and the pressure to attribute specific outcomes to project interventions. Indeed, several attendees said they hoped to learn from other LNC participants how they were connecting interventions with student outcomes.

> I'm eager to try and find out how people are really substantiating the link between professional development and student outcomes. (Education researcher)

We speculate that these two findings—the perceived pressure to produce student learning outcomes and LNC participant expressions of frustration about how to demonstrate project effects on student test scores—may be linked, at least for some MSP participants. We suggest this because some participants appeared to believe that evidence-based outcomes refer exclusively to student outcomes and do not pertain to the outcomes of other types of interventions, such as curriculum development, program improvement, teacher and faculty professional development, and so forth. The following quote suggests this belief:

> The project has not really addressed claims-based outcomes. It has only addressed how the teaching of mathematics can be improved. And I think it has been successful in that, but, you know, it's [ultimate impact on student outcomes is] anyone's guess. (STEM faculty)

We interpret this comment to mean that this respondent did not consider intermediate strategies for improving student learning outcomes a suitable subject of evaluation research that would result in evidence-based outcomes. The view that the only evidence needed is that which pertains to student outcomes may be a function of factors we have previously identified: (a) the national pressure for student outcomes data, (b) a lack of understanding of the complex factors involved in changing education systems, and (c) a lack of experience and training in education research and evaluation.

We have some evidence that respondents with basic insights into the challenges of evaluation alleviated their frustration by adjusting their evaluation goals to focus on measuring the effectiveness of their intermediate intervention strategies. For example, one respondent commented:

> We've been implementing professional development for so many years. Now is the time to evaluate the impact of professional development on student achievement. Are we getting to our goals? Are we doing what we're supposed to do? (Administrator)

IV. RECOMMENDATIONS AND CONCLUSIONS

Based on the data presented in this report, we provide three sets of recommendations: (a) strategies for improving future LNC conferences; (b) strategies for improving project evaluation designs; and (c) strategies for improving the MSP community's expertise in evaluation methodology.

Improving the LNC

In general, the LNC met participants' expectations for developing skills in evaluation design and implementation. Participants expressed a desire to learn "from the field" and acquire evaluation instruments and tools that would enable them to better establish causality between project activities and student achievement. While many respondents expressed satisfaction with the LNC theme and breakout sessions, we believe the following steps may enable future conferences to better serve the MSP community's needs.

Be Cognizant of Problematic Terms Like Claims-Based Outcomes

We found that several participants did not understand the LNC theme of *claims-based outcomes* and expressed different ideas about its meaning and implications for their projects. The term was not well defined or widely accepted, and some viewed it as a slogan or fuzzy catchphrase. This situation is not surprising, given the different disciplines represented in the MSP; some STEM faculty asserted that they had a different and more rigorous understanding of the terms *evidence* and *claims* than other fields. Furthermore, because some participants perceived that the accountability movement (i.e., NCLB) motivated the conference theme, they questioned whether it was merely a fad to wait out or an ongoing priority to be taken very seriously. (We note that participants who are aware of the sustained pressure in the education research community for more rigorous evaluation methods would not be likely to entertain this question.) Given these findings, we recommend that the NSF and LNC organizers select themes or phrases that are either (a) widely known and accepted by the different disciplines active in the MSP or (b) clearly defined in all conference literature and by keynote speakers.

Revise LNC Call for Abstracts to Specifically Focus on Research Design

We found that 79% of the accepted abstracts did not specify a research design type. Moreover, during the conference many participants described challenges pertaining to research design issues. In light of these findings, we recommend that the LNC call for abstracts require much more specific description of the overarching research design or logic guiding a project's research or evaluation activities.

More Explicitly Focus Breakout Sessions on Evidence Generation

Research team observations and some participant comments made clear that about half of the breakout session presenters simply reported on project implementation, rather than relating collected evidence to claims about outcomes. In light of this finding, we recommend that the call for abstracts require not only specific description of research design, but also explanation of the

logical relationship between project data and evidence-based claims.[15] In addition, it might ask writers to identify the stage of their MSP project, so that reviewers can take into account readiness to report evaluation data and make research-based claims. Although projects in their early stages would not be expected to present findings, even early findings, they nonetheless would be expected to describe their evaluation designs and provide information about the instruments and analysis processes they plan to use. This requirement would force these projects to begin the task of developing a strong and feasible design and would provide them with valuable opportunities for feedback and community support.

Consider Using Facilitators for Each Breakout Session

In response to the finding that some breakout sessions were insufficiently interactive or focused on claims-based outcomes, we recommend that each session have a skilled facilitator present. This facilitator should ensure that there is time at the end of the presentations for a Q&A period and that presenters stay on task and focus on evidence-based outcomes topics.

Improving Project Evaluation Designs

The findings presented in this section pertain to projects with inadequate evaluation designs. We believe that further clarifying the role of—and strengthening requirements for—evaluation designs would go a long way toward helping projects that are struggling with this challenge.

Require Proposals to Demonstrate Adequate Evaluation Planning and Budgeting

We found that some participants reflected, in hindsight, that they needed an evaluation logic model with specific measures and instruments or that they wished they could go back and develop a more robust research design prior to the implementation phase. We note that the MSP program solicitation clearly states NSF's expectations regarding project evaluations, including a requirement that proposers describe the "metrics by which partners will document, measure, and report on the project's progress toward realizing improved student and teacher outcomes" (NSF, 2007). We recommend that future MSP program solicitations further require that proposers (a) identify evaluators with strong credentials, (b) provide a preliminary evaluation plan that is guided by the project's overarching logic model and that demonstrates that the evaluator is an integral member of the proposal team, and (c) provide a management plan that makes evident that the evaluator will be integrated into the project leadership structure. Effective interactions with evaluators will help ensure that the evaluation design evolves and remains strong and feasible as the project adjusts in response to obstacles and new opportunities.

It is also important that NSF review panels continue to include reviewers with expertise in evaluation methodology so that they can adequately assess the quality of proposals' evaluation plans. In addition, the evaluation community, especially social scientists who specialize in

[15] While future LNC meetings will have different themes, we anticipate that the focus on evidence-based outcomes will persist in the future.

evaluating education projects, should seek opportunities to serve on such review panels and to become engaged with the MSP community.

Require PIs to Demonstrate Familiarity with Basic Elements of Social Science

The challenges of generating evidence-based outcomes in the domain of education research place special demands on STEM faculty. We recommend that the MSP program take into account that (a) many STEM faculty lack expertise in basic aspects of education research and evaluation design and (b) some STEM faculty appear to believe that evaluation only entails detailed description of project implementation and is not a rigorous scientific endeavor in itself. To overcome these challenges to effective MSP evaluation efforts and bridge the cultural barriers that may exist between STEM faculty and evaluators, we recommend that the MSP program require that PIs demonstrate familiarity with basic elements of social science.

Address Limited Notions of Appropriate Methodologies for Evaluation

Building on the previous point, NSF should also broaden both STEM faculty and evaluator notions of appropriate evaluation methodologies for their projects. While the abstracts submitted to the LNC included a variety of approaches for evaluating project effects (e.g., quasi-experimental designs, social network analysis, mixed-method approaches), the overwhelming majority focused on establishing causal relationships between project activities and student achievement. This finding reflects the national pressure on education reform projects to use methodologies (e.g., RCT and quasi-experimental designs) that are designed to demonstrate statistically significant causal links between specific implementation variables and student outcomes. It also may reflect a tendency for STEM faculty to assume that quantitative approaches emphasizing replicability and generalizability are more rigorous or scientific than other approaches.

While we agree that RCTs and quasi-experimental designs may be the most methodologically rigorous designs available for certain situations, it is evident from respondent comments that such situations are rare in complex MSP projects. Thus, we recommend that the MSP program solicitation encourage proposers to consider diverse evaluation methodologies and select those best suited to the goals and settings of the planned studies. Further encouragement could come in the form of a keynote address at the LNC.[16] The main point would be to encourage members of the MSP community to expand their understanding of designs that are rigorous and appropriate for deriving evidence-based claims about the outcomes of their projects.

Emphasize Value of Measuring Intermediate Effects, Not Just Student Achievement

We learned that (a) most projects are targeting elements of educational systems that are two or more steps removed from actual student learning and that require years before impact on student learning can reasonably be expected and (b) some participants reported that they could

[16] For example, a keynote could address the use of social network analysis (e.g., Spillane et al, 2006) or research-based statistical procedures for measuring a project's effect on professional learning communities or IHE/K–12 networks.

not figure out how to link their upstream activities (such as teacher and faculty workshops) with student achievement. Given this situation, it may be necessary to remind MSP projects of the need to establish annual measurable objectives or benchmarks for each of their goals and to systematically link data over time to summative student outcomes data. For example, projects that are designing and implementing professional development workshops for IHE faculty or K–12 personnel could use pre- and post-measures of workshop participant knowledge gains and changes in classroom practice. This design examines whether, rather than assumes that, there are causal relationships among factors such as teacher content knowledge and student achievement. In other words, we recommend that the MSP program help proposers understand that it is not feasible to produce meaningful findings on student outcomes unless they first focus on assessing effects on their most immediate participants. Developing this understanding is complicated by the absence of evaluation designs that adequately model the nonlinear change processes that take place in complex organizations (Patton, 2006). Identifying such methodologies might be a priority for NSF and the MSP program (see the Improving MSP Community's Expertise in Evaluation Methodology section below).

Require Funded Projects to Establish Procedures for Evaluation Data Sharing

In light of participants' strong desire to learn from their project's evaluation results—whether to make midcourse corrections or correct false assumptions about project efficacy—we recommend that the MSP program continue to establish procedures for project implementers and evaluators to regularly share and discuss evaluation processes and findings. Achieving this objective will require finding ways to navigate disciplinary differences between evaluators and implementers and to establish routines that facilitate information dissemination and collaborative planning across organizational boundaries.

Improving MSP Community's Expertise in Evaluation Methodology

Based on this research, it is clear that many in the MSP community need additional training in education and evaluation research methodologies. This observation pertains not only to STEM faculty, but also to education researchers and professional evaluators, many of whom are also struggling with conducting evaluations in complex organizations. To improve the MSP community's understanding and use of research-based evaluation methods, we make the following recommendations.

Focus Efforts on Training and Professional Development Workshops

Given the challenges that several participants expressed regarding basic evaluation procedures (e.g., having a design in place prior to data collection), it is apparent that the NSF evaluation training strategies currently in place have not yet met the existing need. These strategies include issuing publications about evaluation (e.g., NSF 2005, and myriad other reports and articles) and funding RETA projects designed to provide mentoring in evaluation and other support. Therefore, the MSP program should increase its efforts to provide researchers with the tools, frameworks, and understandings that will allow them to effectively evaluate their projects. We recommend that:

- NSF take an evidence-based approach to this challenge by commissioning a study of the effectiveness of its efforts and then using the findings of that study to design and implement more promising strategies;

- The MSP program hold *mandatory* professional development workshops for PIs and external evaluators at either the LNC or other national meetings of MSP personnel, where participants would work with experts and peers to critique and improve their evaluation and research designs, identify the most appropriate data-gathering instruments, and develop expertise in different analysis methods; and

- NSF develop an indexed repository of research-based methods and evaluation instruments that are particularly relevant to MSPs to help both trained and untrained participants easily identify appropriate evaluation tools.

Explore Methodological Advances for Conducting Evaluations in Complex Institutions

This research found that many participants were struggling to conduct evaluations in complex institutional environments. Some evaluators have argued that the traditional logic model—which posits linear causality between project inputs, outputs, and outcomes—is not sufficient or valid for assessing project effects in complex settings. Relying on a single data type, such as a survey of teacher attitudes about instruction, may also be problematic in providing only a narrow "slice" of data on a factor that is influenced by a complex array of other factors. While a few projects used more robust evaluation designs, including mixed methods, we encourage the MSP program to continue to investigate methodological advances in fields such as evaluation, organizational studies, and education research and to examine ways to communicate this information to proposers and encourage them to consider such methods when designing their evaluations.

In particular, we recommend that the MSP program explore advances in mixed methods that capitalize on the relative strengths of quantitative and qualitative research methods. For example, Osthoff (2007), in assessing the effects of science professional development workshops on K–12 teachers, combined pre- and posttests of content knowledge with in-depth interviews and teacher observations. Taken together, these different methods provide a rich array of data with which to triangulate findings. Triangulation across methods is particularly important if projects are going to assess intermediate effects of their activities, and not simply focus on student achievement as the outcome measure of choice. To enable projects to use these more rigorous approaches, proposers will need examples of methods and theoretical frameworks.

Conclusion

This exploratory research project is an element of the MSP program's own emerging logic model for fostering greater rigor and transparency in evaluation. The program commissioned this study in order to continue its ongoing assessment of (a) the methodologies that MSP grantees are using to make evidence-based claims and (b) the issues and challenges pertaining to evaluation and research that are experienced by the MSP community. As is evident from the findings and recommendations reported here, our research identified factors that must

be addressed in order to advance the MSP community's ability to generate evidence-based claims. Informed by these findings, the MSP program is in a better position to effectively support its grantees as they strive to develop the knowledge and capacity to derive evidence-based claims about the effects of their projects. For this reason alone, we believe that the decision to commission this study places the MSP program in a leadership position within NSF and among other agencies seeking to improve STEM education.

In closing, we emphasize that, although this study raises important questions and illuminates critical issues facing the MSP program, it was not designed to test specific hypotheses or to deeply explore specific topics. Instead, our open-ended approach was intended to sample the breadth of experiences and perceptions among diverse members of the MSP community. While we understand that some of the more specific issues may already be under study by current RETA projects, the MSP program might also consider future research into some of the factors identified here. Such studies might include conducting a survey at a future LNC to gather data on a particular topic (e.g., type and degree of evaluation and research activity focused on identifying the impact of intermediate factors) or undertaking a more complex research effort aimed at testing a specific hypothesis (e.g., projects whose PIs are trained in social science methodology are more likely to use evaluation designs informed by logic models that recognize the nonlinear character of educational institutions).

References

American Evaluation Association. (2003). American Evaluation Association response to U.S. Department of Education notice of proposed priority, Federal Register RIN 1890-ZA00, November 4, 2003. Retrieved April 14, 2009, from http://www.eval.org/doestatement.htm

Bernard, H. R. (2002). *Research methods in anthropology: Qualitative and quantitative approaches* (3rd ed.). Walnut Creek, CA: Altamira Press.

Cernea, M. (Ed.). (1991). *Putting people first: Sociological variables in rural development*. New York: Oxford University Press.

Change and Sustainability in Higher Education. (2006). *Report on course and curriculum changes in Math and Science Partnership (MSP) Programs*. Retrieved April 15, 2009, from http://vipk16.mspnet.org/index.cfm/14133

Frechtling, J. (2007). *Logic modeling methods in program evaluation*. San Francisco: Jossey-Bass.

Katzenmeyer, C., & Lawrenz, F. (2006). National Science Foundation perspectives on the nature of STEM program evaluation. *New Directions for Evaluation, 109,* 7–18.

Kelly, A., & Yin, R. (2007). Strengthening structured abstracts for education research: The need for claim-based structured abstracts. *Educational Researcher, 36*(3), 133–138.

Mosteller, F., Nave, B., & Miech, E. (2004). *Educational Researcher, 33*(1), 29–34.

National Science Foundation. (2005). *Evidence: An essential tool: Planning for and gathering evidence using the Design-Implementation-Outcomes (DIO) cycle of evidence* (NSF 05-31). Arlington, VA: Author. Retrieved April 15, 2009, from http://www.nsf.gov/pubs/2005/ nsf0531/nsf0531.pdf

National Science Foundation. (2007). *Math and Science Partnership (MSP)* (Program solicitation NSF 08-525). Retrieved April 15, 2009, from http://www.nsf.gov/pubs/2008/nsf08525/nsf08525.htm

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Osthoff, E. (2007). *Tentative findings of the SCALE study of grade 6 immersion implementation in LAUSD: A conversation with LAUSD immersion professional development facilitators*. Madison: University of Wisconsin–Madison, Wisconsin Center for Education Research.

Patton, M. (2006, October). *Evidence-based evaluation findings using systems change and complexity science frameworks and ways of thinking*. Paper presented at the MSP Evaluation Summit II, Minneapolis, MN.

Scrimshaw, N. S., & Gleason, G. R. (Eds.). (1992). *Rapid assessment procedures: Qualitative methodologies for planning and evaluation of health related programmes*. Boston: International Nutrition Foundation for Developing Countries.

Spillane, J., Camburn, E., & Lewis, G. (2006). *Taking a distributed perspective in studying school leadership and management: Epistemological and methodological trade-offs*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. Retrieved April 16, 2009, from http://hub.mspnet.org/index.cfm/12853

Strauss, A., & Corbin, J. (1990) *Basic of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Author. Retrieved April 16, 2009, from http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

U.S. Department of Education. (2007). *Report of the Academic Competitiveness Council*. Washington, DC: Author. Retrieved April 16, 2009, from http://www.ed.gov/about/inits/ed/competitiveness/acc-mathscience/report.pdf

W.K. Kellogg Foundation. (2004). *W.K. Kellogg Foundation logic model development guide: Using logic models to bring together planning, evaluation, and action*. Battle Creek, MI: Author. Retrieved April 16, 2009, from http://www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf

Appendix A
Data Collection Instruments

1. Hallway Interview Protocol

A. Consent

Approach a participant, or group of 2 or 3 participants, and obtain consent to interview:

> The information I am collecting will be used to understand the STEM education community's thoughts on evidence-based research. Would you be willing to talk to me for a few minutes? Your participation is voluntary. You may decline to answer questions, and there is no penalty for nonparticipation. There are no risks in participation. Do you consent to being recorded, with the understanding that *we will keep your name, professional affiliation, and MSP affiliation confidential* by not using identifying information in our records?

If they agree to be participate *and* be recorded, turn on recorder:

- State the date and time of day
- Ask:

  > Do you agree to participate in a short audio-recorded interview about the Learning Network Conference, with the understanding that the researchers will keep your name, professional affiliation, and MSP affiliation confidential?

- After they state their agreement, then obtain:
  - Name
  - Professional affiliation
  - MSP affiliation
  - MSP role

If they agree to participate but not to be recorded, proceed with the paper consent form and take handwritten notes.

B. Interview questions

> What are your initial reactions to the conference theme: claims-based outcomes?[17]
>
> Describe your views on the presentations thus far. (Provide examples to illustrate your points)
>
> ➢ If this question was answered in response to question #1, probe for specifics regarding what the participant(s) found useful (or not) in the presentations or talks.
>
> Given your particular MSP role, what has been your experience with claims-based (or evidence-based) outcomes in your MSP project?

[17] The MSP program's working definition of *claims-based outcomes*:
- The phrase *claims-based outcomes* refers to claims or warrants of outcomes of specific interventions or strategies for which there is evidence,
- Where the evidence has been collected using methodologies that would be accepted within the educational research community.
- Hence, this evidence and resulting claims-based outcomes would be recognized and respected in peer-reviewed settings.

- Probe to learn more about the participant's specific angle on the topic, and how it has or has not played out in their own MSP.

C. Interview conclusion

- Provide a sticker and ask the interviewee to wear it for the rest of the conference where it can be easily seen so that other evaluators will not approach them, and
- Say:

  Many thanks for your time and thoughts!

2. Breakout Session Observation Protocol

The goal of the breakout session observations is to capture key aspects of Q&A discussion during and after the main presentations, *particularly as they pertain to the conference theme*. Please focus on the following core topics:

1. Contextual features of the session, including:
   - Numbers attending
   - Room organization and other attributes
   - Mood or demeanor of audience members
2. Types of questions asked, including challenges, supportive statements, clarifications, etc.
   - Include examples of the questions and challenges raised, noting whether any pertain to the claims-based conference theme.
3. Specific points raised during the discussion, especially those that elicited animated responses.
   - Note if points raised pertain to the claims-based theme.
4. Quality of interactions and levels of engagement among participants.
   - Which topics deeply engaged participants?
   - Did the topics that engaged participants pertain to the claims-based theme?
5. Potential connections (conceptual and networking) that people made in the session.
6. Briefly summarize your observations by responding to these questions:
   - Did the presenters stay on task with respect to the conference theme?
   - How did attendees respond, overall, to the presenters, and to each other?
   - To what degree did the audience stay focused on the claims-based conference theme?
   - In your own opinion, how would you assess the quality of the presentations and of the audience response?

3. Think-Piece Exercise on Claims-Based Outcomes

The research group studying the MSP Learning Network Conference is using this 15-minute think-piece activity to learn about the experiences and perspectives of the STEM education community pertaining to claims-based outcomes. Data from all the think-pieces will be analyzed and reported in a report to the NSF and a journal article. Please participate fully in this opportunity to provide the NSF and the wider STEM education community your perspectives by providing as much information as possible.

Please respond to the question that appears at the top of **each** of the attached 3 pages. You may write on the back of a page if you wish.

Before handing in your think piece, please **complete the human subjects consent form** below.

Thank you!

Susan Millar and Matthew Hora
Wisconsin Center for Education Research, UW-Madison

Consent form pertaining to use of your comments in publications resulting from this data gathering activity

Check either A, B, or C:

☐ A. I agree that material from my think piece may be quoted using my name.

☐ B. I agree that material from my think piece may be quoted but without using my name.

☐ C. I do not give permission for material from my think piece to be quoted.

There is no penalty or loss of benefits for nonparticipation; you may end your participation at any time without penalty or loss of benefits. There are no risks in participation. Your responses will be treated confidentially. All data will be held confidential and stored in a secured office, and electronic copies will be stored on a secured computer network.

Signature _____

Name (please print): _____

Primary organizational affiliation: _____

MSP affiliation:_____

MSP role (e.g., evaluator, implementer): _____

[*Each of the following think-piece questions was printed on a separate page in the actual handout used at the conference.*]

Question 1:
Describe your experiences with making evidence-based claims in your MSP, indicating what has and has not been valuable and/or feasible.

Question 2. Pertaining to evaluation:

- For *evaluators*:

a. What factors were/are important in the development and implementation of your MSP's evaluation?
b. How have these factors influenced the evaluation design and its implementation?

- For *implementers*:
  a. How, if at all, have you been involved in your MSP's evaluation?
     i. If you have been involved, what factors affect your implementation of your MSP's evaluation?
  b. Have findings from your MSP's evaluation work been of value to you?

Question 3:

a. How, if at all, have you have used the claims or research of others in your own MSP work?
b. Regarding your responses to 3a., were these claims evidence-based? Whether yes or no, provide examples.

Appendix B
Methodology

Research Design

We based our research design on the rapid assessment approach commonly used in public health and international development program evaluation. This approach employs a mixed-method, interpretive methodology in order to produce a rich, contextualized portrayal of a situation (Scrimshaw & Gleason, 1992). It is particularly useful when time and resources are limited and a multidimensional account of a group's practices, experiences, and attitudes is desired (Cernea, 1991; Bernard, 2002). In undertaking this research, we primarily gathered and analyzed data on participant perceptions and experiences related to the 2-day LNC meeting in January 2008.

Sampling strategies. The sampling universe for this research was all MSP program participants (N = 320), and from this universe different sampling methods were utilized for each type of data collected. First, we used a stratified sample of MSP program participants who submitted presentation abstracts to the LNC to analyze conference abstracts. Second, we used a convenience sample of LNC attendees to select participants for short semistructured interviews. Third, we asked all LNC attendees to write think pieces during the conference.

Data gathering. We collected the following types of data:

- *Presentation abstracts.* Abstracts submitted to the conference planning committee were collected and analyzed, along with limited biographical information about abstract writers. The conference planners solicited the abstracts and biographical information.
- *Semistructured interviews.* Ten trained Westat researchers plus Millar, Hora, Arrigoni, and Kretchmar interviewed a total of 68 conference participants for approximately 10 minutes each, focusing on perceptions, experiences, and opinions about evidence-related issues. Interviews were audiotaped (when background sound level allowed and participants consented) or captured in detailed notes. Millar and Hora developed the interview protocols and met with the Westat researchers before the conference to ensure they understood how the interviews were to be conducted.
- *Breakout session observations.* Guided by an observation protocol, the 10 Westat observers plus Millar, Hora, Arrigoni, and Kretchmar took detailed notes on all 26 breakout sessions. To ensure complete coverage, we asked NSF program officers associated with the MSP program to assist with the observations. Millar and Hora developed the observation protocol, which focused on key topics of interest (e.g., level of experience with evaluation and social science research; constraints on participants' efforts to develop a strong evaluation design). They met ahead of time with the Westat researchers to ensure that they understood the observation methods and for one hour at the end of each day for a debriefing and discussion of emergent findings. All observers organized their observation notes by protocol themes before sending them to Hora. Observers also spent one hour writing up their overall observations of the conference.
- *Think pieces.* All session attendees were asked to produce a short think piece during a 15-minute period prior to the second breakout session on the second day of the meeting. In these think pieces, participants were asked to identify the name of their MSP project and address questions that, in the view of the conference planning committee, would help prepare participants for the breakout session topics while also providing research data. We reviewed

each think piece and identified 98 that included adequate data for further analysis. Our criteria for inclusion in the analysis included detailed answers and evidence of causal or relational thinking (e.g., poor data led to challenges with evaluation). Many of the think pieces included basic descriptions of projects and were not deemed usable for the analysis.

Data Analysis

The four different types of data were each analyzed using analytic methods appropriate for the type, as follows:

- *Presentation abstracts.* To analyze presentation abstract data, we (a) developed a rubric consisting of methodological and activity-based criteria that was based on the LNC call for abstracts; (b) had three staff from the Wisconsin Center for Education Research analyze each abstract using the rubric, after assessing intercoder reliability; and (c) summarized the reviewers' findings for presentation at the LNC and inclusion in the final report.
- *Semistructured interviews, breakout session observations, and think pieces.* Our analytic procedures for this research drew on established procedures of qualitative analysis. These included inductively coding interview transcripts using the grounded theory method of Strauss and Corbin (1990). A coding paradigm—a structured scheme with which to analyze data and identify discrete themes and patterns (Strauss & Corbin, 1990)—is especially needed when working with the kind of multidimensional data collected for this study. The next step of our analysis was centered on reducing our voluminous data set, which involved reviewing highly cited codes, reducing their specificity, and identifying relationships among the most salient and frequently cited factors. To identify these contextual factors, we first looked at the most frequently coded factors. We then ran "reports" on the most frequently cited codes, which listed the interview fragments associated with each code. We reviewed the text and reduced each observation to a shorthand notation, keeping a running list of each observation in a document called a Master Data List (MDL). These observations were then clustered according to theme or type of observation in the MDL, which was used to write the final report.

Appendix C
LNC Call for Abstracts
(http://hub.mspnet.org/index.cfm/14920)

All MSP projects are encouraged to consider presenting key findings on focused aspects of their work during one of the Conference breakout sessions. For this year's Learning Network Conference, the Planning Committee requests that groups interested in presenting at the meeting submit a 3–4 page abstract indicating how the proposed presentation will address the following:

- Context of the work to be presented,
- Claim(s) or hypothesis(es) examined in the work,
- Study design, data collection and analysis,
- Results or knowledge claim, and
- Conclusions and implications.

Abstracts must be related to one of the conference strands (Partnerships, Teaching, Learning). While not limited to the following, topics for presentations that PIs have noted to be of high interest include:

Partnerships

- Determining and measuring the value of partnerships
- Professional learning communities

Teaching

- Increasing teacher content knowledge
- Effectiveness of teacher leaders/coaches
- Mentoring and induction for beginning STEM teachers
- Changing the pedagogy of STEM faculty

Learning

- Measuring student learning outcomes (K–20)
- Serving under-represented student populations

MSP projects are encouraged to submit more than one abstract, discussing and providing evidence for what has been successful and what has not. As the work of the MSPs is ongoing, the research does not have to be fully completed for it to be proposed as a presentation. Completed abstracts must be submitted by Monday, December 3rd to abstracts@mspnet.org using the Conference Abstract Submission Form (click to download MS Word document). Submitted abstracts will be reviewed by the Planning Committee whose members will formally invite presentations by December 21st. Presentations will be invited either for a full 90 minute breakout session or as part of a joint breakout session with another group that has submitted an abstract on the same topic.