# How to Solicit Rigorous Evaluations of Mathematics and Science Partnerships (MSP) Projects

## A User-Friendly Guide for MSP State Coordinators

Coalition for Evidence-Based Policy
A Project Sponsored by The Council for Excellence in Government

NORC
A national organization for research
at the University of Chicago

May 2005

This publication was produced by the Coalition for Evidence-Based Policy, in partnership with the National Opinion Research Center (NORC) at the University of Chicago, under a contract with the Institute of Education Sciences. The Coalition is sponsored by the Council for Excellence in Government (www.excelgov.org/evidence). The views expressed herein are those of the contractor.

This publication was funded by the U.S. Department of Education, Institute of Education Sciences (Contract #ED-01-CO-0028/0001).

# PURPOSE AND OVERVIEW OF THIS GUIDE

**Purpose: To provide MSP state coordinators with a concrete, low-cost strategy to solicit rigorous evaluations of their state's MSP projects.**

Specifically, this strategy will enable states to rigorously answer questions of the type: "Does the MSP project between school district X and college Y increase student math and science achievement and teacher content knowledge; if so, by how much?" The resulting knowledge about "what works" can then be used by schools and districts as an effective, valid tool in ensuring:

(i) that their math and science teachers are highly qualified, and

(ii) that their students are proficient in math and science,

both of which are central goals of American education policy.

The strategy provides MSP state coordinators with a roadmap for soliciting randomized controlled trials (RCTs) – studies which measure a program's impact by randomly assigning individuals (or groups of individuals) to a program group or to a control group. As discussed in the appendix, well-designed RCTs are considered the gold standard for measuring a program's impact, based on persuasive evidence that (i) they are superior to other evaluation methods in estimating a program's true effect; and (ii) the most commonly-used nonrandomized methods often produce erroneous conclusions.

This strategy includes tools that states can use to solicit RCTs of MSP projects that cost as little as $50,000 - $75,000 in some cases, and can produce valid, actionable knowledge about what works within 1-2 years.

**Overview: This Guide provides concrete, step-by-step advice in three areas:**

1. **Overall evaluation strategy:** whether to solicit *single-site* or *cross-site* MSP evaluations.
2. **How to solicit rigorous evaluations:** suggested language for your state's MSP solicitation.
3. **How to (i) review applicants' evaluation plans, and (ii) monitor the evaluations** once underway.

# 1. OVERALL EVALUATION STRATEGY:

## Whether to solicit single-site or cross-site MSP evaluations

### A. Definitions:

- **A single-site evaluation** is an evaluation of a single MSP project, to determine its effectiveness.

- **A cross-site evaluation** is the evaluation of multiple MSP projects that are implementing a specific, well-defined MSP model (e.g., the Chicago Math and Science Initiative, or the Milken Teacher Advancement Program). Such an evaluation addresses the question, "how effective is this particular MSP model as implemented in a range of MSP projects." To carry out such an evaluation, you would need to ensure that all sites in the evaluation implement the same MSP model (using solicitation language such as that discussed on page 11).

### B. Key factors to consider in deciding whether to solicit single-site versus cross-site evaluations.

- **A cross-site evaluation may be appropriate if you have strong reason to believe that a particular MSP model will be effective in a range of MSP sites.**

  A rigorous cross-site evaluation, by assessing the model's effectiveness in different school districts, with different students and teachers, will likely yield strong evidence to confirm or disprove your preliminary judgment. If the evaluation finds that the model is indeed effective across different sites, you and others would then have a strong basis for replicating the model at other MSP sites across the state or national MSP program. Such a finding would represent an important development for American math and science education – a field where very few interventions have been proven effective when implemented across different sites.

  One cautionary note: many examples exist of highly-promising educational interventions that, when subjected to a rigorous cross-site evaluation, were found marginally effective or ineffective.

- **Soliciting a few single-site evaluations may be appropriate if you wish to encourage a diversity of MSP approaches in your state, then rigorously determine which are effective.**

  This approach – encouraging local experimentation coupled with rigorous evaluation – may be the preferred route when you do not have strong preliminary evidence supporting a specific MSP model. A rigorous single-site evaluation will generate strong evidence about the effectiveness of an MSP approach as implemented in one site. Subsequent cross-site evaluations would then be needed to determine whether the MSP approach is effective in different settings.

- **Some MSP projects, however, may not have enough math/science teachers to meet the sample needed for a rigorous single-site evaluation (i.e., about 60 teachers).**

  Specifically, for an MSP evaluation (single-site or cross-site) to produce strong evidence about an MSP project's effect on student math or science achievement, a minimum sample of about 60 teachers (plus their classes) is needed – 30 in the program group and 30 in the control group. This estimate assumes that the true effect of the MSP project on student achievement is modest in size (e.g., increases math achievement in grades 1-5 by at least 25 percent of a grade level per year).[1] If the true effect of the MSP project on student math or science achievement is large, a smaller sample – e.g., 20 teachers plus their students – may suffice.[2] But, if at all possible, we would urge a minimum sample of 60 teachers, for reasons discussed in the endnote.[3]

  These estimates of the minimum sample size assume that the MSP project provides roughly the same professional development program to all participating teachers (e.g., the same summer training course provided to all participating middle school math teachers). If instead the MSP project provides different programs to different teachers (e.g., one summer course for math teachers, another for science teachers), then the minimum sample is about 60 teachers *per program*.

  Many individual MSP projects have enough math and science teachers to meet these sample size requirements, in which case a rigorous single-site evaluation is feasible. Some local MSP projects, however, may not have enough teachers. If you wish to rigorously evaluate the effectiveness of these smaller projects, you will need to solicit a cross-site evaluation.

# 2. HOW TO SOLICIT RIGOROUS EVALUATIONS:

## Suggested language for your state's MSP solicitation

This section contains step-by-step guidance on soliciting rigorous MSP evaluations, including suggested solicitation language (shown in the shaded boxes). Guidance on soliciting single-site evaluations is immediately below; guidance on soliciting cross-site evaluations starts on page 11.

### A. Guidance on soliciting single-site evaluations.

The following solicitation provisions are designed to solicit rigorous single-site evaluations.

**Solicitation provision 1 – to incentivize MSP grant applicants to build a rigorous evaluation into their projects.**

MSP applicants are encouraged to build a high-quality randomized controlled trial (RCT) into the design of their project, in order to rigorously evaluate its effectiveness. RCTs are considered the gold standard for measuring a project's impact, based on persuasive evidence that (i) they are superior to other evaluation methods in producing valid estimates of a project's impact; and (ii) the most commonly-used nonrandomized methods often produce erroneous conclusions. Applicants that propose an RCT in their Evaluation Plan will receive:

- [*Fill in number*] additional points in the proposal review process [*e.g., 15 additional points out of a possible 100*].
- A grant supplement of $50,000 to $75,000 to help cover the cost of the evaluation, assuming the project is selected for award.

Applicants will receive the additional points and grant supplement if their proposed RCT (including the proposed research team) is judged by reviewers to be of high quality. Small MSP applicants which, by themselves, may have not have the required minimum sample of teachers to carry out an RCT, can also receive the additional points and grant supplement if they propose to partner with other MSP applicants to carry out a cross-site RCT. Applicants partnering in this way would need to implement the same MSP model (e.g., the same summer institute program providing the same teacher training).

Applicants can download a User-Friendly Guide to high-quality, low-cost RCTs in the MSP program from the U.S. Education Department's web site (http://www.ed.gov/programs/mathsci/resources.html).

The above web site contains an electronic copy of this Guide.

Provision 1 might also offer applicants a smaller number of additional points (e.g., 5) if, instead of an RCT, they propose a high-quality matched comparison-group study (see discussion of such studies in the appendix).

You could provide the grant supplement of $50,000-$75,000 to applicants proposing a high-quality RCT by (i) proportionately reducing the size of the grants awarded to other applicants; or (ii) contributing some of the state-level funding used to administer the MSP program.

**Solicitation provision 2** – **to request the applicant's proposed plan for carrying out the RCT.**

We suggest that you include this provision in your solicitation's section on Evaluation Plans:

Applicants that propose to evaluate their project in an RCT, per section [ ] of this solicitation [*fill in section containing provision 1*], should include the following items in their Evaluation Plan:

**A.** **A short statement of the research question that the RCT seeks to answer** (e.g., "Does the MSP project increase student math achievement; if so, by how much?")

**B.** **Identification of a researcher, or research team, who (i) has agreed to carry out the RCT, and (ii) who has previous experience in carrying out a high-quality RCT.**

It may not be necessary for the lead researcher to have previous experience in carrying out an RCT as long as a key member of, or consultant to, the research team has such experience. Please attach a copy of a previous RCT that the researcher or other experienced team member has carried out.

**C.** **A brief description of the plan, developed by the applicant and researcher, for recruiting the required sample of teachers to participate in the RCT.**

Minimum sample size requirements are discussed on page 5 of the U.S. Education Department's User-Friendly Guide to RCTs in the MSP program, at http://www.ed.gov/programs/mathsci/resources.html.

The applicant's plan for recruiting teachers into the study should:

- Provide assurance that the participating school district(s) have agreed to the random assignment process; and
- Describe what steps the study will take to recruit the required sample of teachers.

**D. Brief assurances that the applicant and researcher will ensure the integrity of the randomization through the following steps:**

- Having someone independent of the MSP project carry out the lottery or other process for random assignment of teachers.
- Asking teachers in the intervention group not to share MSP program materials with teachers in the control group (so as to avoid contamination of the control group).
- Ensuring that the schools' assignment of teachers to their *classes* is unaffected by whether the teachers are in the intervention or control group. (If a school assigns teachers to classes based on who participates in the MSP training – e.g., gives the intervention group teachers the advanced classes – it will undermine the equivalence of the intervention and control groups.) Describe briefly what steps the applicant and researcher will take to ensure this, such as (i) assigning teachers to their classes *prior to* randomizing teachers to the intervention and control groups; (ii) keeping the school principal or other person who assigns teachers to classes unaware of which teachers are in the intervention and control groups; or (iii) randomly assigning teachers to their classes.
- Collecting and analyzing outcome data for all teachers randomly assigned to the intervention and control groups, even those intervention-group teachers who do not actually complete the MSP intervention. (This is known as an "intention-to-treat" approach, and is designed to ensure that the intervention and control groups remain equivalent over the course of the study – i.e., have no systematic differences other than those caused by the intervention.)
- Making every effort to obtain outcome data for at least 80 percent of the teachers originally randomized, and the students entering their classes. (This is known as "maximizing sample retention" and is designed to ensure that the intervention and control groups remain equivalent over the course of the study.) As part of such assurance, describe briefly the steps the applicant and researcher will take to maximize sample retention, such as obtaining test scores for students in the study who transfer to another school within the same district or state.

**E. A brief description of how the study will measure project outcomes.**

- The study should use standardized tests of student math and/or science achievement as one of the outcome measures, since a key goal of the MSP program is to increase student achievement by enhancing the knowledge and skills of their teachers. If feasible, the study should also use teacher content knowledge as an outcome measure.

**Explanation of item E above**: Measuring project outcomes.

- **We suggest student achievement as an outcome measure not only because increasing it is a key program goal, but also because it can often be measured at low cost.**

  Indeed, the overall cost of the RCT may be as low as $50,000 to $75,000 if the RCT measures outcomes using achievement test scores that schools *already* collect for other purposes. Such a low cost is possible because an RCT's largest cost is usually collecting

the outcome data. For many MSP projects, it is now possible to carry out a low-cost RCT to evaluate the project's impact on student math scores, because many states now test mathematics achievement annually, especially in the early grades. Testing of students' science achievement is less common, so one possible strategy for low-cost evaluation of MSP projects is to assess their impact on student math, but not science, achievement.

To measure MSP project outcomes using existing achievement tests, one other condition must apply – namely, the researcher must be able to obtain test scores for individual students, not just aggregate grade-level or school-level test scores. This is because the researcher will need to compare test scores of the students in the program group to those of students in the control group.

- **As an additional option, you may wish to use teacher content knowledge as an outcome measure.**

  This is because improving such knowledge is a key intermediate goal of the MSP program. However, measuring impact on teacher content knowledge poses two additional challenges. First, it requires that tests be administered to teachers in the study, which raises the study's cost. Second, the minimum sample of teachers that we suggested for the study – 60 teachers (see page 5) – may not be large enough to generate strong evidence about the effect on teacher content knowledge (for reasons, see endnote 3). Instead, a minimum sample of 90 teachers is probably needed to generate such evidence.[4] Often, individual MSP projects may not have enough math or science teachers to meet this sample size requirement.

  If the individual MSP projects in your state do not have enough teachers to meet this minimum sample size, yet you still wish to measure their impact on teacher content knowledge, we suggest two possible courses of action. First, you could solicit a cross-site evaluation to obtain the required sample, rather than single-site evaluations (see pages 4-5). Alternatively, you could go ahead and use teacher content knowledge as an outcome measure anyway, recognizing that such an approach may show effects on teacher content knowledge that do not reach statistical significance and therefore constitute *suggestive* evidence, rather than strong evidence. Such suggestive evidence is useful in generating hypotheses to test in larger evaluations in the future.

- **In addition, if resources permit, we suggest measuring (i) project implementation, and (ii) long-term educational outcomes.**

  Illustrative project implementation measures include the extent to which (a) MSP trainers cover the key items in the MSP curriculum, (b) participating teachers complete the MSP training sessions, and (c) the teachers apply new skills learned in the MSP sessions to their classrooms; and other such measures of implementation fidelity and follow-through. Measuring implementation can help generate the information needed to replicate the MSP project should it prove to be effective, or illuminate possible reasons why it is not effective should that be the study's finding.

  Illustrative long-term educational outcomes include: (a) student test scores over a several-year period, (b) special education placements and grade retentions, (c) student

enrollment in higher-level math and science courses, and (d) high-school graduation rates. Measuring such outcomes would enable you to assess whether the MSP project has a sustained impact on student success or failure in school over time.

**Solicitation provision 3 – to request periodic reports on the evaluation, once underway.**

We suggest that you include this provision in your solicitation's section on Reporting Requirements. The provision requests grantees, on a semi-annual basis, to complete a short checklist of items that are critical to the study's success. A significant departure from any one of these may indicate that a further expenditure of funds on the evaluation is unwarranted.

Applicants that are evaluating their project in an RCT, per section [ ] of this solicitation [*fill in section containing provision 1*], must submit a brief (e.g., 1-2 page) semi-annual report on the study's implementation, addressing and/or updating the following checklist of items. These checklist items are critical to the successful implementation of the study; a significant departure from any one of these may well undermine the validity of the study's conclusions:

- ☐ Was the lottery or other process for random assignment of teachers carried out by someone independent of the MSP project? Who?
- ☐ Did the number of teachers randomized to the intervention and control groups meet or exceed the minimum sample size (as described in the Guide to RCTs in the MSP program, at http://www.ed.gov/programs/mathsci/resources.html)? Please provide a complete roster of teachers assigned to the intervention and control groups.
- ☐ Did the evaluator ensure that the schools' assignment of teachers to their *classes* was unaffected by whether the teachers were in the intervention or control group? (A school's assignment of teachers to classes based on who participates in the MSP training would undermine the randomization.)
- ☐ Of those teachers randomized to the intervention group, did any drop out of the MSP project, and is the researcher still collecting and analyzing outcome data for these teachers? (It is important that the researcher do so, to maintain the integrity of the randomization.)
- ☐ Did any teachers in the control group, or their students, cross over to the intervention group after random assignment? (Such crossovers undermine the randomization.)
- ☐ For what percentage of the teachers and students in the RCT did evaluators obtain outcome data? (To maintain the integrity of the randomization, it is important for the researcher to obtain such data for at least 75-80 percent of the teachers originally randomized, and the students entering their classes – the higher the better.)
- ☐ Does the researcher's analysis of study outcomes use tests for statistical significance that are based on both the number of teachers randomized and the number of their students, rather than just on the total number of students in the study? (Such "hierarchical" analysis is critical to obtaining valid tests of statistical significance.)

## B. Guidance on soliciting cross-site evaluations.

**Step 1** – **Issue a solicitation for a researcher to carry out the evaluation.**

We suggest that this solicitation contain provisions 2 and 3 above, requesting the research applicant's proposed evaluation plan (including research team), and semi-annual reports on the evaluation once underway. These provisions will need to be slightly modified, as follows:

- Adjust both provisions to reflect the fact that they are addressed to research applicants rather than MSP applicants.
- Delete the first bullet in provision 2C, requesting applicants to assure that school districts have already agreed to the random assignment process.
- Add the following item at the end of provision 2, to ensure that all sites in the study implement the same MSP model:

> F. A brief description of the applicant's plan to ensure that all the sites in the study implement the same MSP model and adhere closely to the model's specific features.

**Step 2** – **Include a provision in your state's MSP grant solicitation to incentivize MSP applicants to participate in the cross-site evaluation.**

We suggest a provision such as the following:

> The state has enlisted [*insert name of researcher selected in step 1*] to carry out a rigorous cross-site evaluation of a particularly promising MSP model – the [*insert name of MSP model*]. The evaluation will include an experimental design involving the random assignment of teachers to program and control groups, so as to generate scientifically-valid evidence on the effectiveness of this MSP model.
>
> To facilitate the evaluation, MSP applicants that have an agreement with [*insert name of researcher*] to (i) implement the [*name of MSP model*], and (ii) participate in the cross-site evaluation, including the random assignment of teachers, will receive [*fill in number*] additional points in the proposal review process [*e.g., 15 additional points out of a possible 100*]. Please provide a copy of communications (e.g., letter or email exchange) showing that such an agreement is in place.

In this provision, you may also wish to (i) describe the MSP model being evaluated; and (ii) discuss any steps the study will take to address applicants' concerns about random assignment. Such steps might include offering control-group teachers an alternative program of professional development, or offering them participation in the MSP project after a one-year or two-year delay if it proves effective (see appendix for further discussion).

# 3. HOW TO (I) REVIEW APPLICANTS' EVALUATION PLANS AND (II) MONITOR THE EVALUATIONS ONCE UNDERWAY

## A. Suggested criteria for reviewing an applicant's evaluation plan (either single-site or cross-site).

- **Does the proposed research team include a researcher, or key consultant, who has previously carried out a well-designed RCT?**

  We suggest that you use this as a central review criterion, since a successful RCT generally requires the involvement of a researcher (or key consultant) with hands-on experience and demonstrated success in carrying out such a study. If, as suggested above, your solicitation has asked the applicant to attach a previous RCT that the researcher or consultant has carried out, we recommend that you review it using *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide,* a publication of the U.S. Education Department's Institute of Education Sciences (at http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf). Appendix B of that publication contains a clear, one-page checklist you can use to review the quality of the researcher's previous study.

- **Does the applicant have a sound, workable plan for carrying out the evaluation?**

  That is, does the applicant's evaluation plan effectively address the key items requested in solicitation provision 2 (above), including (i) how the applicant will recruit the required sample of teachers, (ii) ensure the integrity of the randomization, and (iii) measure MSP project outcomes?

## B. Suggestions for monitoring an evaluation once it is underway.

If your solicitation has asked awardees to send you brief semi-annual reports addressing the checklist of key items in solicitation provision 3 (above), it should be straightforward for you to monitor the evaluation over time. The seven items on this checklist are critical to the success of the study. If the awardee departs significantly from any one of these, you may wish to consider discontinuing funding for the study.

# APPENDIX: WHY THIS GUIDE SUGGESTS RANDOMIZED CONTROLLED TRIALS (RCTS) TO MEASURE MSP PROJECTS' EFFECTIVENESS

Well-designed RCTs are considered the gold standard for measuring an intervention's impact across many diverse fields of human inquiry, such as education, welfare and employment, medicine, and psychology.[5] This is based on persuasive evidence that (i) they are superior to other methods in estimating an intervention's true effect; and (ii) the most common study designs – including "pre-post" studies and "comparison-group" (or "quasi-experimental") studies without careful matching – often produce erroneous conclusions. Thus, this Guide suggests that MSP state coordinators solicit RCTs, where feasible. The following discussion elaborates, and also suggests an alternative when RCTs are not feasible.

## A. Definition: RCTs are studies that measure an intervention's effect by randomly assigning individuals (or groups of individuals) to an intervention group or a control group.

For example, suppose that a school district wants to rigorously evaluate whether a new teacher professional development curriculum is more effective than the district's existing curriculum. The district might undertake an RCT which randomly assigns teachers to either an intervention group, which receives the new curriculum, or to a control group, which uses the existing curriculum. The RCT would then measure outcomes – such as teacher content knowledge or test scores of their students – for both groups over a period of time. The difference in outcomes between the two groups would represent the effect of the new curriculum compared to the existing curriculum.

## B. The unique advantage of random assignment: It enables you to assess whether the intervention itself, as opposed to other factors, causes the observed outcomes.

Specifically, the process of randomly assigning a sufficiently large number of individuals into either an intervention group or a control group ensures, to a high degree of confidence, that there are no systematic differences between the groups in any characteristics (observed and unobserved) except one – namely, the intervention group participates in the intervention, and the control group does not. Therefore, assuming the RCT is properly carried out, the resulting difference in outcomes between the two groups can confidently be attributed to the intervention and not to other factors.

## C. Evidence supporting RCTs: There is persuasive evidence that –

(i) Well-designed RCTs are superior to other study designs in estimating an intervention's true effect; and

(ii) Well-matched comparison-group designs may be a good alternative when an RCT is not feasible.

**Specifically:**

- **"Pre-post" study designs often produce erroneous results.**

  **Definition**: A "pre-post" study examines whether participants in an intervention improve or become worse off during the course of the intervention, and then attributes any such improvement or deterioration to the intervention.

  The problem with this type of study is that, without reference to a control group, it cannot answer whether the participants' improvement or deterioration would have occurred anyway, even without the intervention. This often leads to erroneous conclusions about the effectiveness of the intervention.

  **Example.** A pre-post study of Even Start – a federal program designed to improve the literacy of disadvantaged families – found that the children in the program made substantial improvements in school readiness during the course of the program (e.g., an increase in their national percentile ranking on the Picture Peabody Vocabulary Test from the 9th to the 19th percentile). However, an RCT of Even Start carried out by the same researchers found that the children in the *control* group improved by approximately the same amount over the same time period. Thus, the program had no *net* impact on the children's school readiness. If the researchers had only carried out the pre-post study, and not the RCT, their results would have suggested erroneously that Even Start is highly effective in increasing school readiness.[6]

- **The most common "comparison group" study designs (also known as "quasi-experimental" designs) also lead to erroneous conclusions in many cases.**

  **Definition:** A "comparison group" study compares outcomes for intervention participants with outcomes for a comparison group chosen through methods other than randomization.

  For example, a comparison-group study of a new teacher professional development curriculum might compare outcomes for teachers who receive the new curriculum to outcomes for a group of teachers in a neighboring school who do not receive the new curriculum.

  In education and other areas, a number of "design replication" studies have been carried out to examine whether and under what circumstances comparison-group studies can replicate the results of RCTs. These investigations have shown that most comparison-group studies in education and other areas of social policy produce inaccurate estimates of an intervention's effects. This is because of differences between the intervention and comparison groups that differentially affect their outcomes.[7]

- **<u>However,</u> *<u>well-matched</u>* <u>comparison-group studies</u> can produce valuable knowledge, and may be a good alternative when an RCT is not feasible.**

  Specifically, the design replication studies noted above generally support the value of comparison-group studies in which the comparison group is *very closely matched* with the intervention group –e.g., in student test scores prior to the intervention, demographic characteristics, time period in which the two groups are studied, and methods used to collect their outcome data. Among comparison-group studies, these well-matched studies are the most likely to generate valid conclusions about an intervention's effectiveness. However, their estimates of the magnitude of an intervention's effect are often inaccurate, and in some instances they still produce erroneous overall conclusions about whether the intervention is effective, ineffective, or harmful.

  This body of evidence therefore suggests that well-matched comparison-group studies can establish *possible* evidence of an intervention's effectiveness, thereby generating good hypotheses that merit confirmation in RCTs. And in cases where RCTs are not feasible or not yet available, such well-matched studies may serve as a second-best alternative.

## D. RCTs may not be feasible in some cases – e.g., due to study participants' concerns about random assignment.

For example, in the MSP program, some schools and/or teachers may have concerns about randomly assigning some teachers to a control group that will not participate in the MSP project. We believe there are often effective strategies that you can use to address and overcome their concerns (discussed immediately below); however, if these are unsuccessful, you may wish to solicit well-matched comparison-group studies as a second-best alternative. If you do, we suggest you keep in mind that very careful matching of the intervention and comparison group – particularly on student test scores prior to the program – increases the chances that the study will produce valid estimates of an MSP project's effect.

## E. You may be able to overcome schools' and teachers' concerns about random assignment through steps such as the following:

- In cases where an MSP project cannot enroll all eligible teachers due to budget or capacity limitations, you can make a strong case that random assignment – i.e., a lottery – is a fair way to determine which teachers will participate.
- As discussed in this Guide, you can offer a competitive priority and/or larger award amount to MSP grant applicants that agree to participate in an RCT.
- You can offer control-group teachers participation in the MSP project after a one-year or two-year delay, if the project proves to be effective.

- You can offer control-group teachers an alternative program of professional development. The RCT would then be evaluating the effectiveness of the MSP project *compared to* that of the other program.

# Notes

[1] This estimate of sample size is based on the following assumptions: The desired power for the study is 0.80; the project's true effect size is at least 0.2 standard deviations (i.e., in the modest range); each teacher has 35 students total in his or her math/science classes; the intra-class correlation is 0.075; a covariate (baseline test scores) with a 0.8 correlation with outcomes is used in estimating the project's effect; the study seeks to estimate the project's effect at the .05 level of significance in a two-tailed test; and the study obtains outcome data for 80% of the original sample of teachers and students.

[2] This estimate of sample size is based on the same assumptions as in endnote 1, with one modification: the project's true effect size is large – at least 0.4 standard deviations (e.g., it increases math achievement in grades 1-5 by at least half a grade level per year).

[3] We would urge a minimum sample of 60 teachers, if possible, for two reasons. Most importantly, as noted in the main text, it will enable the evaluation to identify MSP projects that have a modest or large effect on student achievement – not just those with a large effect. But in addition, it increases the evaluation's ability to identify MSP projects that improve teacher content knowledge, as well as student achievement. Measuring the effect on teacher content knowledge generally requires a larger sample than that needed to measure the effect on student achievement because, in measuring teacher content knowledge, one loses the statistical power that comes from including teachers *with their classes of students* in the sample.

[4] This estimate of sample size is based on the following assumptions: The desired power for the study is 0.80; the project's true effect on teacher content knowledge is at least 0.4 standard deviations; a covariate (baseline test scores) with a 0.8 correlation with outcomes is used in estimating the project's effect; the study seeks to estimate the project's effect at the .05 level of significance in a two-tailed test; and the study obtains outcome data for 80% of the original sample of teachers.

[5] See, for example, Office of Management and Budget, *Program Assessment Rating Tool (PART) Guidance for FY 2006 Budget*, p. 24, http://www.whitehouse.gov/omb/part/2006_part_guidance.pdf; the Food and Drug Administration's standard for assessing the effectiveness of pharmaceutical drugs and medical devices, at 21 C.F.R. §314.12; "The Urgent Need to Improve Health Care Quality," Consensus statement of the Institute of Medicine National Roundtable on Health Care Quality, *Journal of the American Medical Association*, vol. 280, no. 11, September 16, 1998, p. 1003; and *Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination*, Society for Prevention Research, April 12, 2004, at http://www.preventionresearch.org/sofetext.php.

[6] Robert G. St. Pierre et. al., "Improving Family Literacy: Findings From the National Even Start Evaluation," Abt Associates, September 1996.

[7] Howard S. Bloom et. al., "Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?" MDRC Working Paper on Research Methodology, June 2002, at http://www.mdrc.org/ResearchMethodologyPprs.htm. James J. Heckman et. al., "Characterizing Selection Bias Using Experimental Data," *Econometrica*, vol. 66, no. 5, September 1998, pp. 1017-1098. Daniel Friedlander and Philip K. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *American Economic Review*, vol. 85, no. 4, September 1995, pp. 923-937. Thomas Fraker and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *Journal of Human Resources*, vol. 22, no. 2, spring 1987, pp. 194-227. Robert J. LaLonde, "Evaluating the Econometric Evaluations of Training Programs With Experimental Data," *American Economic Review*, vol. 176, no. 4, September 1986, pp. 604-620. Roberto Agodini and Mark Dynarski, "Are Experiments the Only Option? A Look at Dropout Prevention Programs," Mathematica Policy Research, Inc., August 2001, at http://www.mathematica-mpr.com/PDFs/redirect.asp?strSite=experonly.pdf. Elizabeth Ty Wilde and Rob Hollister, "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Scores as Outcomes," Institute for Research on Poverty Discussion paper, no. 1242-02, 2002, at http://www.ssc.wisc.edu/irp/.

This literature is systematically reviewed in Steve Glazerman, Dan M. Levy, and David Myers, "Nonexperimental Replications of Social Experiments: A Systematic Review," Mathematica Policy Research discussion paper, no. 8813-300, September 2002. The portion of this review addressing labor market interventions is published in "Nonexperimental versus Experimental Estimates of Earnings Impact," *The American Annals of Political and Social Science*, vol. 589, September 2003.