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

A User-Friendly Guide for MSP Project Officials and Evaluators

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

August 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 (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).

This publication is in the public domain. Authorization to reproduce it in whole of in part for educational purposes is granted.

PURPOSE AND OVERVIEW OF THIS GUIDE

Purpose: To provide MSP project officials and evaluators with clear, practical advice on how to conduct rigorous evaluations of MSP projects at low cost.

Specifically, this is a how-to Guide designed to enable MSP grantees and evaluators of MSP projects to answer questions about the projects' impact, such as: "Does this MSP project improve student math and science achievement and increase 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 advice in this Guide is targeted primarily to two audiences:

(i) MSP grantees (or grant applicants) that are working with an evaluator to conduct a rigorous single-site evaluation (i.e., evaluation of the grantee's single MSP project); and

(ii) Evaluators that have been selected by a state to conduct a rigorous cross-site evaluation (i.e., evaluation of multiple MSP projects that are implementing a specific, well-defined MSP model, such as the Chicago Math and Science Initiative).

We have also prepared a companion Guide for MSP state coordinators – *How to Solicit Rigorous Evaluations of MSP Projects* – at http://www.ed.gov/programs/mathsci/issuebrief.doc.

Our advice is designed to enable MSP projects and evaluators to conduct rigorous impact evaluations that (i) are low in cost (e.g., $50,000 - $75,000 for a single-site evaluation), and (ii) produce valid, actionable knowledge about what works within 1-2 years.

Overview: This Guide provides concrete, step-by-step advice, as follows:

Step 1: Find a researcher with expertise in conducting rigorous impact evaluations to include on the study team.

Step 2: Decide what research question(s) the study seeks to answer.

Step 3: Decide on the study design.

Step 4: Gain the cooperation of teachers and school officials.

Step 5: Allocate teachers to the program and control (or matched comparison) groups.

Step 6: Collect the data needed to measure the MSP project's effectiveness.

Step 7: Analyze and report the study's results.

STEP 1: FIND A RESEARCHER WITH EXPERTISE IN CONDUCTING RIGOROUS EVALUATIONS TO INCLUDE ON THE STUDY TEAM.

This may be the most important step in the process. This Guide provides general principles in conducting these evaluations; however, a researcher with hands-on experience and demonstrated success in conducting such studies is needed to address specific operational questions as they arise over the course of the study. Such a researcher might participate as the principal investigator or other study team member, or as a key consultant to the team.

To identify such a researcher, we suggest that:

A. **You contact the authors of previous well-designed impact evaluations** to explore their interest in participating in the study and/or their recommendations of colleagues to do so. To find such authors, you can go to websites that summarize findings from well-designed randomized controlled trials (the strongest study design) and well-matched comparison-group studies, such as the sites listed in Appendix A. These sites all identify the study authors – including many capable mid-level researchers who might participate on your study team at modest cost. In some cases, the sites also list the authors' contact information.

Another useful place to look for such researchers is the What Works Clearinghouse's Registry of Outcome Evaluators, at http://whatworks.ed.gov/technicalassistance/EvlSearch.asp. Keep in mind that the evaluators on this list are self-nominated; therefore, a careful review is needed to identify those who may have previously conducted a well-designed impact evaluation.

B. **You ask the prospective researchers that you so identify for a plan that describes, in clear, nontechnical language, how they would carry out Steps 2-7 in this Guide.**

We suggest you review the plan to determine whether it:

- Asks clear research questions about the MSP project's effect on student achievement and, if appropriate, teacher content knowledge (Step 2);
- Includes a strong study design – preferably a randomized controlled trial or, if not possible, a well-matched comparison-group study (Step 3A);
- Includes a sound, workable approach to recruiting the minimum sample of teachers (Step 3B), gaining the cooperation of teachers and school officials (Step 4), and allocating teachers to program and control (or matched comparison) groups (Step 5);
- Includes outcome measures (e.g., student achievement scores) that will enable you to answer the study's research questions (Step 3D), and a sound, workable approach to collecting outcome and other data (Step 6); and
- Includes a sound, workable approach to analyzing the study results, so as to obtain valid estimates of the MSP project's effects on key outcomes (Step 7).

C. **You ask the prospective researchers for references**, such as other researchers or school officials who participated in the researcher's previous studies.

We suggest you ask the references for evidence that the researcher: (i) has played a central role in an earlier rigorous evaluation; (ii) successfully handled that role; and (if the researcher is being considered for principal investigator) (iii) has the organizational and interpersonal skills to manage the operation of a study, staying within budget and schedule.

STEP 2: DECIDE WHAT RESEARCH QUESTION(S) THE STUDY SEEKS TO ANSWER.

We suggest you choose just a few well-defined research questions – ones which (i) will yield valuable, actionable knowledge about "what works" in the MSP program, and (ii) can preferably be answered at low or modest cost. More specifically:

A. **We suggest that the study seek to evaluate a specific, well-defined MSP approach.**

For example, suppose that your MSP project(s) provides one training program to elementary school teachers, and a different training program – with a distinct curriculum – to middle school teachers. In this case, we suggest that your study evaluate one of the training programs or the other, but not both (unless resources permit you to conduct a study of each, with the appropriate sample size of teachers in each study). This is because if you evaluate the two together, you may be able to determine whether the two together are effective, but you will probably not be able to determine, with statistical confidence, whether it was one program or the other or both that produced the study's overall finding. You may also not be able to find common outcome measures for the two different programs – a problem which may well prevent the study from producing meaningful results.

B. **We suggest that one of the research questions address the effect of the MSP project(s) on student math and/or science achievement**, since (i) a key goal of the MSP program is to increase student achievement by enhancing the knowledge and skills of their teachers, and (ii) this effect can often be measured at very low cost, using test scores that schools *already* collect for other purposes.

Specifically, you can often measure the effect on student *math* achievement at low cost because many states now test mathematics achievement annually, especially in the early grades. Testing of students' *science* achievement, however, is less common. Thus one possible low-cost approach would be to choose a research question about the project(s) effect on math, but not science, achievement (and address science achievement if additional resources become available for that purpose). For example:

> "*Does the MSP project between school district X and college Y increase student math achievement in grades 2-5, compared to a control group? If so, by how much?*"

The overall cost of a single-site RCT addressing this central policy question may be as low as $50,000-$75,000, if the study measures achievement using test scores that schools already collect.

C. An additional research question might be the effect of the MSP project(s) on teacher content knowledge, since improving such knowledge is a key intermediate goal of the MSP program. For example:

> "*Does the MSP project increase the math and/or science content knowledge of middle-school science teachers, compared to a control group? If so, by how much?*"

However, measuring the effect on teacher content knowledge poses two additional challenges:

- It requires a larger sample of teachers to identify an effective project, for reasons discussed in the companion Guide for MSP state coordinators on soliciting rigorous MSP evaluations (http://www.ed.gov/programs/mathsci/issuebrief.doc, page 9). Suggestions for addressing the sample size issue are discussed in Step 3B, below.

- It requires that a survey or other instrument be administered to teachers to assess their content knowledge, which raises the study's cost.

D. If resources permit, you may also wish to choose research questions about the longer-term effect of the MSP project(s).

These might include, for example, questions about whether the effect of teachers' MSP training on their students' achievement increases or decreases in successive school years – e.g., because teachers need time to fully incorporate their new knowledge into classroom instruction, or alternatively because over time they forget what they learned. Thus, you might include research questions about the effect of teachers' MSP training in the summer of 2006 on the achievement of their 2006-2007 students, and on the achievement of their 2007-2008 students.

You might also ask research questions about the effect on students' educational performance over a 2-3 year period (e.g., math/science test scores, enrollment in higher-level math/science courses, grade retentions and special education placements).

E. If resources permit, you might also choose one or two research questions relating to MSP project implementation. For example:

> "*How many hours of training did teachers in the MSP project receive on average, and what percentage of teachers completed all MSP training sessions?*"
>
> "*To what extent did MSP trainers cover the key items in the MSP curriculum?*"
>
> *"What non-MSP professional development programs do teachers in the control group participate in, what training curriculum is used, how many hours of such training do they receive on average?*"

Such research questions can help you interpret the study findings, identifying (i) possible reasons why the project is effective or ineffective, and (ii) if effective, likely ingredients needed to replicate it successfully in other sites.

STEP 3: DECIDE ON THE STUDY DESIGN.

This includes the following sub-steps: (A) Deciding whether to use a randomized controlled trial or well-matched comparison-group study design; (B) Deciding on the sample size (i.e., number of teachers to include in the study); (C) Deciding how teachers will be recruited into the study, and allocated between program and control (or matched comparison) groups; and (D) Deciding how to measure project outcomes.

A. Deciding on overall design: We strongly suggest a randomized controlled trial (RCT), and would recommend a well-matched comparison-group design only if an RCT is truly infeasible.

If at all possible, we suggest that you conduct an RCT, in which teachers (plus their classes) are randomly assigned to the MSP project or to a control group. As discussed in the companion Guide for MSP state coordinators (http://www.ed.gov/programs/mathsci/issuebrief.doc, pages 13-16), 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; (ii) the most commonly-used nonrandomized methods often produce erroneous conclusions and can lead to practices that are ineffective or harmful.

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. In most cases, a persistent research team can assuage their concerns and gain their cooperation in the RCT using approaches such as those discussed in Step 4 of this Guide.

We suggest that you consider a well-matched comparison-group study *only* if you have exhausted all possible options for conducting an RCT and conclude that it is truly not feasible. A well-matched comparison-group study compares outcomes for teachers (plus their classes) that participate in the MSP project against outcomes for a comparison group of teachers that (i) are chosen through methods other than randomization, and (ii) are closely matched with the MSP teachers in their students' initial achievement levels and other characteristics (as discussed further in Step 3C).

Among nonrandomized studies, comparison-groups studies with such careful matching are the most likely to generate valid estimates of an intervention's true effect, but we suggest them only with reservation because they may still produce erroneous conclusions (as discussed in the companion Guide, http://www.ed.gov/programs/mathsci/issuebrief.doc, page 15).

B. Deciding on sample size:

1. To measure the effect on student achievement, a minimum sample of about 60 teachers, plus their classes, is needed.

As discussed in the companion Guide for MSP state coordinators (http://www.ed.gov/programs/mathsci/issuebrief.doc, page 5), a minimum sample of about 60 teachers is needed 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. Of

the 60, 30 teachers plus their classes would be randomly assigned to the MSP program group, and 30 would be randomly assigned to the control group. (Similarly, in a matched comparison-group study, 30 would be included in the program group and 30 in the comparison group).[1]

Many individual MSP projects have enough math and/or science teachers to meet these sample size requirements, in which case a rigorous single-site evaluation is feasible. However, some smaller MSP projects may not, by themselves, have enough teachers. Such projects may therefore wish to partner with other MSP projects to carry out a cross-site evaluation with at least 60 teachers. Projects partnering in this way would need to implement the same MSP model (e.g., the same summer institute program providing the same teacher training), for reasons discussed above (Step 2A).

2. **To measure the effect on teacher content knowledge, a larger sample of teachers is needed (e.g., 90).**

   The reasons for the larger sample requirement are discussed in the companion Guide for MSP state coordinators (http://www.ed.gov/programs/mathsci/issuebrief.doc, page 9). If your MSP project does not have enough teachers to meet this requirement, yet you still wish to measure its effect on teacher content knowledge, we suggest two possible courses of action.

   First, you could partner with other MSP projects to carry out a cross-site evaluation with at least 90 teachers. Alternatively, you could go ahead and measure the effect on teacher content knowledge anyway, using a smaller sample (e.g., 60 teachers). It is important to recognize, however, that such an approach (i) is likely to show statistically-significant effects on teacher content knowledge only for MSP projects that are highly effective, and (ii) in other cases, may show effects 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.

C. Deciding how to recruit and allocate teachers:

1. **The evaluator should preferably recruit teachers into the study using the same process the MSP project would use in the absence of a study.**

   This will help ensure that the evaluation is measuring the effectiveness of the MSP project as it is normally implemented. So, for example, if the MSP project would normally recruit teachers by publicizing its training program and asking teachers to volunteer, the study should preferably do the same. Alternatively, if the MSP project would normally ask school principals to designate teachers to participate, the study should recruit in the same way.

   In a cross-site evaluation (i.e., evaluation of several MSP projects all implementing the same MSP model), we suggest you recruit teachers from each MSP project roughly in proportion to the size of the project.

2. **The simplest way to allocate teachers in an RCT: Apply random assignment to the entire sample, ensuring equal numbers in the program and control groups.**

For example, if you have recruited 60 teachers from several schools to participate in the study, the simplest approach would be to randomly assign 30 teachers to the MSP project and 30 to a control group, using computer software that ensures an equal number in each group. The random assignment will ensure, to a high degree of confidence, that there are no systematic differences between the two groups of teachers and their classes in initial student achievement levels and other characteristics.

If, however, the teachers' classes have *widely* varying initial levels of math and/or science achievement (e.g., for reasons discussed in endnote 1), it is possible that applying the above random assignment process to a sample of just 60 teachers may not produce two equivalent groups. In such circumstances, you may wish to consider "blocked" random assignment. Under this approach, you would group teachers into several "blocks" (e.g., those teaching high-achieving students, those teaching average-achieving students, and those teaching low-achieving students), and then randomly assign teachers *within each block* to program and control groups. As an illustrative example, you might group the whole sample of 60 teachers into three blocks of 20, and then randomly assign 10 teachers in each block to the program group and 10 to the control group.

In cases where teachers' classes vary widely in initial achievement, blocking may increase the study's ability to produce precise estimates of an MSP project's effects. (In other cases, it may actually do the reverse.) Therefore, you may wish to ask the member of your research team with expertise in RCTs whether blocking makes sense for your study.[2]

3. **<u>In a comparison group study</u>, select comparison-group teachers who are *very closely matched* with program group teachers**, particularly in their classes' initial achievement level.

   We offer here a few general principles on matching, and suggest that you consult a researcher with expertise in such studies for further advice:

   - <u>Of primary importance</u>, the comparison-group teachers should be very closely matched with program group teachers in their classes' initial math and/or science achievement level (whichever will be the main outcome of interest).

   - <u>Of secondary importance</u>, the teachers in the program and comparison group should also be matched in the demographics and grade levels of their students, and geographical proximity (e.g., same school district).

   Statistical techniques such as propensity score matching may be used to accomplish such matching, but a full discussion of such techniques is beyond the scope of this Guide.[3]

   Careful studies have shown that when a comparison-group study does not include close matching in the above characteristics, the study is unlikely to generate accurate results even when statistical methods (such as regression adjustment) are used to correct for these differences in estimating the program's effect.

   Finally, the comparison group should not be comprised of teachers who had the option to participate in the MSP project but declined. This is because such teachers may differ systematically from the teachers who do participate in their level of motivation and other

important characteristics that are not easily measured. The difference in motivation (or other characteristics) may itself lead to different outcomes for the two groups, resulting in inaccurate (and potentially misleading) estimates of the MSP project's true effect.

D. <u>Deciding how to measure MSP project outcomes</u>:

1. **<u>For student achievement</u>, outcomes should preferably be measured with test scores on established, standardized tests that the schools already administer.**

   Such test scores have credibility in the eyes of school officials, and because schools already collect them for other purposes, using these scores to measure MSP study outcomes can greatly reduce the cost and administrative burden of the MSP study (as noted in Step 2).

   To use such data to measure MSP project outcomes, three conditions must apply:

   - **First, the evaluator must be able to obtain test scores for individual students**, not just aggregate grade-level or school-level test scores. This is because the evaluator will need to compare test scores of the students being taught by program-group teachers to the test scores of students taught by control-group teachers. (Suggestions for gaining access to individual test scores are discussed in Step 4.)

   - **Second, the evaluator must be able to obtain students' test scores both before they enter the classes of teachers in the study, and at the end of the study.** Their pre-program scores will be used as "covariates" in estimating the MSP project's effect on student achievement, as discussed in Step 7.

   - **Third, the evaluator must be able to convert the scores to a form that will enable a comparison of student achievement across grade levels** – such as a district, state, or national percentile ranking, or to categories such as "advanced," "proficient," or "basic." Illustrative examples of studies that do this include Rouse and Krueger's RCT of Fast ForWord (which converted the state's standardized reading assessment for students in grades 3-6 to a district percentile ranking)[4] and the New York City Voucher Experiment (which converted scores on the Iowa Test of Basic Skills for students in grades 4-7 to a national percentile ranking).[5]

2. **<u>To measure teacher content knowledge</u>, evaluators should preferably use a standardized instrument** whose ability to accurately measure such knowledge is well-established, as opposed to an instrument developed by the evaluator or MSP project for the purposes of the study. The instrument should be capable of measuring the particular areas of teacher content knowledge that the MSP training seeks to enhance. Several standardized instruments to measure teacher math and science content knowledge currently exist and may fit this purpose.[6]

STEP 4: GAIN THE COOPERATION OF TEACHERS AND SCHOOL OFFICIALS.

For an evaluator to conduct an RCT, teachers and school officials will need to agree to the random assignment of teachers (plus their classes) to program and control groups. In addition, in either an RCT or matched comparison-group study, school and/or district officials will need to provide the evaluator with test scores for the individual students in the teachers' classes (as discussed in Step 3D). Teachers and school officials may raise ethical concerns about either or both of these actions – concerns which should be fully aired and addressed at the start of the study. What follows are suggestions to MSP project officials and evaluators for addressing such concerns, in order to gain the teachers' and school officials' cooperation in these key elements of the study.

A. A key first step toward gaining their cooperation is to identify one or more senior-level advocates for your study at the district and/or state level.

In a single-site evaluation of an individual MSP project, the superintendent of the participating school district has presumably agreed to the study, and thus may provide senior-level support for the study. Also, in either a single-site or cross-site evaluation, the MSP state coordinator who is soliciting the evaluation may serve this role. In addition, you can often find other senior district or state officials who are proponents of education research generally (e.g., Directors of Research, Evaluation, Improvement), and who might therefore be enlisted to support your study.

As discussed below, such senior-level advocates can be extremely helpful in gaining the cooperation of teachers and school officials – by participating in your meetings with them and underscoring the study's importance; by interceding with school or district staff whose excessive caution or overly-narrow legal interpretations are thwarting your ability to conduct the study; and/or by vouching for you as a trustworthy, credible researcher.

B. Specific suggestions for gaining teachers' and school officials' cooperation in random assignment.

Teachers and/or school officials sometimes object to random assignment on the grounds that it is unethical to deny a beneficial program – in this case, the MSP training - to half the teachers in the study, when such training could significantly improve their teaching ability and thus their students' education. To address this issue, we suggest that you do the following.

1. **Meet with the school officials and teachers whose cooperation you seek. Include your senior-level advocate(s), to underscore the study's importance.**

   You may wish to meet first with the school officials (e.g., principals) whose cooperation you seek. Then, once you have enlisted their support, you may wish to meet with the teachers.

   In the meetings, we suggest that you describe the MSP project and the study to the teachers and/or school officials, and give them an opportunity to fully discuss any issues or concerns they may have.

2. **Make points such as the following to gain their cooperation in random assignment:**

   - If, as is often the case, the MSP project cannot enroll all eligible teachers due to budget or capacity limitations, random assignment – i.e., a lottery – is arguably the fairest way to determine which teachers will participate.

     It may be unethical not to evaluate the MSP project with the strongest possible methodology – i.e., random assignment. Nonrandomized methods have been shown, in some cases, to produce erroneous conclusions and lead to practices that are ineffective or harmful (see companion Guide for MSP state coordinators, pp. 14-15). To continue a program that is ineffective is a disservice to teachers and students, as well as the taxpayers who fund it.

   - This study will produce scientifically-valid, actionable results that the school and/or district can then use to meet key educational goals: (i) ensuring that their math and science teachers are highly qualified, and (ii) ensuring that their students are proficient in math and science.

   - This is an important study that could influence state and national policy on teacher professional development.

3. **Where appropriate, explore possible modifications to the study design, such as the following, to address their concerns:**

   - Offer control-group teachers participation in the MSP project after a one-year or two-year delay, if the project proves to be effective.

   - Offer school officials a limited number of discretionary exemptions from random assignment – e.g., the opportunity to designate one or two teachers who will not be subject to random assignment but instead will be guaranteed participation in the MSP project. You would not include these teachers in the study sample.

C. **Specific suggestions for gaining access to the individual test scores of students in the MSP study.**

1. **MSP evaluators working with a district or state can legally gain access to such scores, with appropriate precautions to protect student privacy.**

   Specifically, the Family Educational Rights and Privacy Act (FERPA) allows "organizations conducting studies for, or on behalf of, educational agencies … for the purpose of … improving instruction" to gain access to individual student educational records, including test scores, without the consent of their parents "if such studies are conducted in such a manner as will not permit the personal identification of students and their parents by persons other than representatives of such organizations and such information will be destroyed when no longer needed for the purpose for which it is conducted." [Title 20 U.S.C., Section 1232(g)(b)(1)(F)]

2. **To reassure school and/or district officials that you are a trustworthy evaluator who will faithfully observe such privacy precautions,** we suggest steps such as the following:

- Offer to sign a written agreement (i) stating that you will abide by the confidentiality provisions in the law, and (ii) recognizing that unauthorized disclosure of student educational records is illegal. As an illustrative example, we have attached the standard form for such an agreement used by the Austin, Texas Independent School District (see Appendix B).

- Ensure that your evaluation team includes someone who is trusted and respected by school and district officials.

- Provide the names of education officials whom you worked with in previous studies, who can vouch for you as a trustworthy partner.

3. **Enlist the help of your senior-level advocate(s) at the district or state level** to persuade officials to provide you with the test score data.

D. Special advice on gaining school and teacher cooperation, for <u>evaluators selected by the state to conduct a cross-site evaluation</u> of multiple MSP projects.

Whereas the evaluator in a *single-site* MSP evaluation presumably has already partnered with a school district to evaluate the district's MSP project, the evaluator conducting a *cross-site* evaluation for the state faces the challenge of recruiting MSP projects to (i) participate in the study, and (ii) implement the specific MSP model that the state seeks to evaluate (e.g., the Milken Teacher Advancement Program).

We suggest that cross-site evaluators, in addition to taking the steps above to gain school and teacher cooperation, do the following:

- **Enlist the active assistance of <u>state education officials</u> in recruiting MSP projects.** For example, if the state's MSP solicitation provides a competitive preference for MSP grant applicants that agree to participate in the cross-site evaluation (as suggested in the companion Guide for MSP state coordinators, http://www.ed.gov/programs/mathsci/issuebrief.doc, page 11), state officials can help greatly by (i) publicizing that preference to grant applicants during the solicitation period, (ii) referring applicants that express interest in participating in the study to the evaluator; and (iii) after grant award, enforcing grantees' agreement to participate in the evaluation.

- **Enlist the active assistance of the <u>developer of the MSP model</u> being evaluated.** The developer presumably can make a strong case for the benefits of its MSP model, and therefore can be a valuable ally in recruiting schools to participate in the study of that model. Also, during the study, the developer should play the lead role in ensuring that the MSP projects in the study are implementing their MSP model effectively (e.g., adhering closely to the model's training curriculum).

E. Special advice on gaining school and teacher cooperation, for <u>evaluators using a matched comparison group of teachers from another school district</u> not participating in the MSP project.

Such evaluators face the challenge of persuading a school district whose only role in the study is to supply the comparison-group teachers and students – and which therefore may not have a

strong interest in the study – to allow the evaluator to collect student test scores and other data for members of the comparison group (as discussed in Step 6). Here are a few suggestions for such evaluators in gaining the cooperation of the district and its schools:

- Reassure the district and/or school officials that you will faithfully protect the confidentiality of student test score data (see Step 4C, above).
- Point out that they, as a potential participant in this MSP project in the future, will benefit from learning whether the project is effective or not, and offer to provide them a special briefing on the results at the end of the study.
- If appropriate, offer to provide the MSP training in their district after the study ends, if the study shows it is effective.
- Keep the process of collecting data on the comparison-group teachers and their students as short and streamlined as possible.
- Enlist the assistance of your senior-level advocate(s) at the state level in gaining the cooperation of the school district (e.g., by reinforcing the points above, and emphasizing the study's importance).

STEP 5: ALLOCATE TEACHERS TO THE PROGRAM AND CONTROL (OR MATCHED COMPARISON) GROUPS.

A. In an RCT, the evaluator should conduct the random assignment of teachers to the program and control groups, rather than MSP project officials.

(In a matched comparison-group study, the evaluator should construct the comparison group, rather than MSP project officials.)

This is to ensure that the process is carried out by persons without a vested interest in the study outcome. In an RCT, the evaluator should conduct the random assignment using computer software that protects against possible manipulation of the process, and ensures that an equal number of teachers are assigned to the program versus control group.[7] (As discussed in Step 3, if random assignment is "blocked," the software should ensure that an equal number of teachers are assigned to the program versus control group within each block.)

B. In an RCT, random assignment should preferably be carried out as close in time to the start of the MSP training as possible.

(In a matched comparison-group study, the construction of program and comparison groups should be done as close in time to the start of the MSP training as possible.)

This will help minimize the number of "no-shows" – i.e., teachers in the program group who don't participate in the MSP training because, for example, they leave their job or switch jobs before the training begins.

C. **Evaluators and MSP project officials should ask teachers assigned to the program group not to share their MSP training materials with teachers in the control group** (or matched comparison group). Such sharing would "contaminate" the control (or comparison) group, and thus undermine the study's ability to estimate outcomes for the program group *compared to* what their outcomes would have been without the MSP training.

D. **The evaluator should ensure that schools' assignment of teachers to their *classes* is unaffected by whether the teachers are in the MSP program group or control/comparison group.**

This is because if a school assigns teachers to classes based on who participates in the MSP training (e.g., gives the MSP program group teachers the advanced classes), it will undermine the equivalence of the program and control groups (e.g., cause the program group to be disproportionately comprised of advanced classes, and thus possibly lead to an erroneous finding that the MSP program increases student achievement).

We suggest three possible ways to ensure schools' assignment of teachers to classes is unaffected by their participation in the MSP training: (i) The school could assign teachers to their classes *prior to* the randomization of teachers to the program and control groups; (ii) the school principal or other person who assigns teachers to classes could be kept unaware of which teachers are in the MSP program group versus the control group; or (iii) teachers in the study could be *randomly* assigned to their classes.

STEP 6: COLLECT THE DATA NEEDED TO MEASURE THE MSP PROJECT'S EFFECTIVENESS.

A. **As a general matter, we suggest you keep the data collection process as short and streamlined as possible**, limiting it to data that is truly needed to answer the research questions in Step 2. Keeping it short and streamlined will help in gaining the cooperation of teachers and schools officials, and in obtaining a high response rate.

B. **At the time of random assignment (or construction of program and matched comparison groups), evaluators should collect the following brief data on all study participants:**

- **A unique personal identifier for each teacher and each of his or her math/science students**, so that the status and outcomes of all teachers and their students can be accurately tracked over the course of the study. Be sure never to include these identifiers in any study reports or other releases in a way that might enable readers to trace data to specific students. (Doing so is a violation of Family Educational Rights and Privacy Act.)

- **For each teacher: (i) whether he or she has been assigned to the program or control group, and (ii) the identity of each student in his or her classes.**

- **Brief demographic data on each teacher and student** (e.g., for teachers, years of education, degree received, years of teaching experience; for students, age, race, sex, free or reduced-price lunch status). We suggest you obtain as much of this data as possible from school records, so as to minimize the need to administer a survey to the teachers and students.

C. <u>At the time of random assignment</u> (or construction of program/comparison groups) evaluators should also obtain data on the <u>pre-program</u> performance of study participants.

More specifically, evaluators should obtain pre-program measures of the outcomes that the MSP project seeks to improve. For example:

- If student test scores on the state's standardized math achievement test will be used as an outcome measure, evaluators should obtain such scores for students around the time they enter the classes of the teachers in the study. (Alternatively, evaluators could simply verify with school officials that such pre-program scores exist for these students, and then actually collect the scores at the same time they collect the students' post-program test scores, at the end of the school year.)

- Similarly, if teacher content knowledge will be used as an outcome measure, evaluators should administer their instrument for measuring such knowledge to all teachers in the study just prior to random assignment. (If they administer it after random assignment, they may have trouble persuading some teachers in the control group to take the test.)

D. <u>During the MSP training and subsequent school year</u>, evaluators should collect the data needed to answer their research questions (if any) about project implementation (questions such as those discussed in Step 2).

These data might include, for example:

- Number of training sessions each teacher in the MSP project attends, and length of each session;

- The extent to which MSP trainers cover the key items in the MSP training curriculum; and

- For each teacher in the program and control (or matched comparison) group, whether he or she participates in non-MSP professional development activities, number of hours of such participation, and curriculum used in these activities.

Such data can be obtained through brief surveys of the teachers and/or trainers, evaluator observations of the training sessions, session attendance lists, and similar sources.

E. <u>At the end of the school year</u>, evaluators should obtain the outcome data on student achievement and (if applicable) teacher content knowledge for study participants.

Step 3D (above) offers advice on what outcome data to collect. We suggest collecting outcome data on teacher content knowledge at the end of the school year, rather than immediately

after the MSP training, in order to determine whether the training produced more than a temporary improvement in such knowledge.

F. **<u>After the end of the MSP project</u>, evaluators should obtain the data needed to answer their research questions (if any) about the longer-term effect of the MSP project** (questions such as those discussed in Step 2).

These data might include, for example, the test scores of the students that they teach in succeeding school years – e.g., two or three school years after end of the MSP training. Such data would be needed to determine whether the effect of the teachers' MSP training on their students' achievement increases or decreases in successive school years.

G. **Evaluators should make every effort to obtain outcome and other data for at least 80% of the original sample of teachers, and the students who entered their classes.**

This effort ("maximizing sample retention") is critical to ensuring that the program and control (or matched comparison) groups remain equivalent over the course of the study. The most challenging part of this effort is usually obtaining outcome data for students who transfer to another school during the course of the study. For students who transfer to another school *within the same district*, evaluators should be able to obtain their test score data with the assistance of district officials. For students who transfer to *another district or state*, obtaining such data is more difficult; however, the number of such transfers may well be minimal during the course of the study.

The following are a few suggestions for maximizing sample retention during the course of the study:

- Periodically provide a roster of teachers in the study to school officials, and ask if any of the teachers have left their teaching position; and

- Periodically contact the teachers in the study to ask if any of their students have transferred out of their classes.

Most schools maintain forwarding information for students and teachers who transfer, enabling you to follow up to obtain their outcome data if it appears sample retention may fall below 80 percent.

H. **Evaluators should seek outcome data for *all* teachers (plus their students) in the original program and control (or matched comparison) groups** -- even those program-group teachers who do not actually enroll in or complete the MSP training ("no-shows"), and even those students who transfer out of their math or science class during the school year.

This known as an "intention-to-treat" approach, and is designed to ensure that the program and control groups remain equivalent over the course of the study. A common mistake of program evaluators is not to seek outcome data for the no-shows, or use these data in estimating the program's effect – a mistake which may well result in inaccurate estimates of the program's effect.[8]

STEP 7: ANALYZE AND REPORT THE STUDY'S RESULTS.[9]

A. Evaluators should obtain regression-adjusted estimates of the MSP project's effect on student achievement, using available software programs for analyzing the results of group-randomized trials.[10] We suggest regression-adjusted impact estimates, rather than simple impact estimates, because adjusting for students' pre-program test scores with a regression covariate (as discussed below) is likely to greatly increase the precision of your estimates.

The regressions should:

1. **Use tests for statistical significance that are based on both the number of teachers in the study 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. A common mistake of program evaluators conducting studies in which groups are randomized (in this case, teachers plus their classes of students) is to conduct tests for statistical significance using only the total number of group members (i.e., number of students). This mistake may well lead to an erroneous finding that the program's effect is statistically significant.[11]

2. **Use the students' pre-program test scores as a covariate.** Doing so may greatly increase the study's ability to produce a finding that the MSP project's effects are statistically significant.

3. **Estimate the MSP project's overall effect on student achievement, and preferably its effect on various subgroups as well** (e.g., students with low levels of initial achievement, low-income students as measured by reduced-price lunch status, minority students).

   It is often feasible, in studies such as this which randomize teachers plus their classes, to analyze the project's effect on subgroups of *individual students* (e.g., students with low initial test scores) but not subgroups of *classes* (e.g., 7th grade classes in Monroe Middle School). We therefore do not recommend that you conduct subgroup analyses for classes.[12]

4. **Use outcome data for *all* teachers and their students in the original sample for whom such data is available,** even teachers who do not complete the MSP training and students who transfer to other math/science classes during the school year (consistent with the "intention-to-treat" approach discussed in Step 6).

In addition, if the evaluator "blocked" the random assignment (as discussed in Step 3C), the regressions should include a separate dummy variable for each block except one.

B. Evaluators, in reporting these estimates of the effect on student achievement, should:

1. **Indicate whether they are statistically significant at conventional levels (usually .05), and**

2. **Preferably translate them into grade-level equivalents**, so that they can be easily understood by nonresearchers.

   Translating the effects into grade-level equivalents can usually be accomplished by ascertaining, in the regressions above, the average increase in achievement of the control-group students over the course of the school year, and then dividing the project's estimated effect by that amount.[13]

C. Evaluators should obtain and report regression-adjusted estimates of the MSP project's effect on <u>teacher content knowledge</u> (if used as an outcome measure).

This analysis will be similar to that above for student achievement, except that (i) the tests for statistical significance should be based on the number teachers only (rather the number of teachers and their students); and (ii) the regressions should use pre-program measures of the teachers' content knowledge as a covariate (rather than students' pre-program test scores).

D. Evaluators should obtain and report regression-adjusted estimates of the MSP project's effect on any <u>long-term outcomes</u> measured (such as those discussed in Step 6F).

E. Evaluators should summarize and report data they obtain on <u>project implementation</u> (per Step 6D) in clear, accessible terms.

F. Evaluators should briefly summarize and report <u>descriptive statistics</u> on the teachers and students in the study, collected per Step 6B (e.g., teachers' years of education and teaching experience; students' age, race, sex, free or reduced-price lunch status).

These descriptive statistics should include the average <u>pre-program</u> student test scores and (if measured) teacher content knowledge for (i) the overall sample; (ii) the program group; and (iii) the control (or matched comparison) group. As a check on whether random assignment (or matching) produced two equivalent groups, evaluators should test and report on whether the two groups differed in any of these pre-program characteristics.

G. The evaluators should conduct and report an <u>analysis of the study's attrition</u> (i.e., inability to collect outcome data for all teachers and students originally randomized).

1. **This should include the following statistics on the study sample:**

   - The number of teachers originally allocated to the program and control (or matched comparison) groups;
   - The number of the students initially enrolled in their classes;
   - The number of such teachers and students in (i) the program group and (ii) the control (or matched comparison) group, for whom outcome data was obtained.[14]

2. **In addition, evaluators should conduct and report tests of whether study attrition caused differences between the program and control (or matched comparison) groups.**

These should include statistical tests to determine if there are significant differences in pre-program test scores between (i) those students lost to follow up in the program group, and (ii) those students lost to follow up in the control (or matched comparison) group.

H. Evaluators should <u>summarize the study's design and implementation</u> in a concise discussion of the study's methodology.

APPENDIX A: WEBSITES LISTING RIGOROUS EVALUATIONS, WHICH YOU CAN USE TO IDENTIFY CAPABLE RESEARCHERS FOR YOUR STUDY TEAM

What follows is a reproduction of Appendix A of the Institute of Education Sciences' publication *Identifying and Implementing Educational Interventions Supported By Rigorous Evidence: A User-Friendly Guide*, at http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html. It lists websites that summarize the results of well-designed RCTs and matched comparison-group studies. These sites all identify the study authors, including many capable mid-level researchers who might participate on your study team at modest cost. Some of these sites also provide the authors' contact information, to facilitate your following up with them.

Where to find evidence-based interventions

The following web sites can be useful in finding evidence-based educational interventions. These sites use varying criteria for determining which interventions are supported by evidence, but all distinguish between randomized controlled trials and other types of supporting evidence. We recommend that, in navigating these web sites, you use this Guide to help you make independent judgments about whether the listed interventions are supported by "strong" evidence, "possible" evidence, or neither.

The What Works Clearinghouse (http://www.w-w-c.org) established by the U.S. Department of Education's Institute of Education Sciences to provide educators, policymakers, and the public with a central, independent, and trusted source of scientific evidence of what works in education.

The Promising Practices Network (http://www.promisingpractices.net/) web site highlights programs and practices that credible research indicates are effective in improving outcomes for children, youth, and families.

Blueprints for Violence Prevention (http://www.colorado.edu/cspv/blueprints/index.html) is a national violence prevention initiative to identify programs that are effective in reducing adolescent violent crime, aggression, delinquency, and substance abuse.

The International Campbell Collaboration (http://www.campbellcollaboration.org/) offers a registry of systematic reviews of evidence on the effects of interventions in the social, behavioral, and educational arenas.

Social Programs That Work (http://www.evidencebasedprograms.org) offers a series of papers developed by the Coalition for Evidence-Based Policy on social programs that are backed by rigorous evidence of effectiveness.

APPENDIX B: ILLUSTRATIVE EXAMPLE OF AGREEMENT ASSURING SCHOOL DISTRICT YOU WILL PROTECT CONFIDENTIAL STUDENT DATA

Access to Confidential Data
AISD Staff/Non-AISD Partner Agreement

Please sign and return with the Application for External Research and Evaluation.

Director of Research
or Evaluation Project: _____
Title of Project: _____

Address: _____

City/State/Zip: _____

Telephone: Work_____ Home_____

Fax_____ E-mail_____

Justification: Educational Research and Evaluation

I understand that any unauthorized disclosure of confidential information is illegal as provided in the federal Family Educational Rights and Privacy Act of 1974 (FERPA), 20 U.S.C. 1232 *eg. seq.* and in the implementing federal regulations found in 34 CFR Part 99. FERPA is specifically incorporated into the Texas Public Information Act (formerly known as the Open Records Act). It is listed as an exception to records that are subject to disclosures to the public.

In addition, I understand that any data, datasets or output reports that I, or any authorized representative, may generate are confidential and the data are to be protected. I will not distribute to any unauthorized person any data or reports that I have access to or may generate using confidential data.

I hereby agree that failure to abide by the requirements of this client agreement may lead to the immediate revocation of any contract (or research study) that I may be performing for AISD. I understand that any intentional, knowing, or negligent release of confidential student information to unauthorized persons may also subject me to a legal cause of action for violation of an individual's civil rights in addition to state or federal criminal penalties.

_____
Director of Research or Evaluation Project's Signature Date

_____
Sponsor of Research or Evaluation Project's Signature Date

Notes

[1] You may need a larger sample than 60 teachers if the teachers' classes vary widely in their initial levels of math and/or science achievement. The may be the case, for example, if there is a high degree of ability grouping between classes; if the teachers' classes are comprised of very different types of students (e.g., 6th graders performing in the lowest 25th national percentile, and 10th graders performing in the top 25th national percentile); or if some teachers are in high-poverty, low-achieving schools and others are in low-poverty, high-achieving schools. If such conditions apply in your MSP project, we suggest you consult with a researcher or statistician with expertise in this issue ("power analysis in group randomized trials") to determine the minimum sample for your study.

[2]For a helpful discussion of blocked random assignment, see Howard S. Bloom, *Randomizing Groups To Evaluate Place-Based Programs*, March 2, 2004, on the William T. Grant Foundation web site (http://www.wtgrantfoundation.org/usr_doc/RSChapter4Final.pdf)

[3] There exists an extensive literature on propensity score matching and other nonrandomized methods, addressing the extent to which they can produce valid estimates of a program's effect. 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, http://www.mathematica-mpr.com/publications/PDFs/nonexperimentalreps.pdf.

Other important contributions to this literature include the following. 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/. 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.irp.wisc.edu/.

[4] Rouse, Cecilia Elena and Alan B. Krueger, "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically-based' Reading Program," *Economics of Education Review,* volume 23, issue 4, August 2004, p. 323.

[5] Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell, *School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program*, Mathematica Policy Research, February 19, 2002, at http://www.mathematica-mpr.com/publications/pdfs/nycfull.pdf.

[6] Examples of existing instruments for measuring teacher content knowledge include: Praxis II (http://www.ets.org/praxis/prxva.html); instruments developed by the Learning Mathematics for Teaching project (http://www.soe.umich.edu/lmt/); and Surveys of Enacted Curriculum (http://www.ccsso.org/projects/Surveys_of_Enacted_Curriculum/).

[7] For helpful advice on how to implement random assignment, see Orr, Larry L., *Social Experimentation: Evaluating Public Programs With Experimental Methods: Implementation and Data Collection*, prepared for the Assistant Secretary of HHS for Planning and Evaluation, October 1997, at http://aspe.hhs.gov/search/hsp/qeval/part5.pdf.

[8] For a useful discussion of why an "intention-to-treat" approach is needed to obtain valid estimates of a program's effect, see Institute of Education Sciences, *Identifying and Implementing Educational*

Interventions Supported By Rigorous Evidence: A User-Friendly Guide, at http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html, page 7. The intention-to-treat results can often be used to calculate the program's effect on those program-group members who actually complete the program, through a "no-show" adjustment – see Orr, Larry L., *Social Experimentation: Evaluating Public Programs With Experimental Methods: Basic Concepts and Principles of Social Experimentation,"* prepared for the Assistant Secretary of HHS for Planning and Evaluation, October 1997, at http://aspe.hhs.gov/search/hsp/qeval/part2.pdf

[9] Illustrative examples of study reports for randomized controlled trials can be found on the website for the Institute of Education Sciences' National Center for Education Evaluation, at www.ed.gov/about/offices/list/ies/ncee/index.html. Please keep in mind that these are much larger studies, and more extensive reports, than those envisioned in this Guide.

[10] Such software can also be used to obtain regression-adjusted estimates for matched comparison-group studies in which groups (as opposed to individuals) are matched. For helpful advice on obtaining impact estimates in group-randomized trials, see Howard S. Bloom, op. cit., no. 2. For helpful advice on obtaining regression-adjusted impact estimates generally, see Orr, Larry L., *Social Experimentation: Evaluating Public Programs With Experimental Methods: Analysis,* prepared for the Assistant Secretary of HHS for Planning and Evaluation, October 1997, at http://aspe.hhs.gov/search/hsp/qeval/part6.pdf.

[11] Specifically, using the total number of students in the study as the sample size will cause you to understate the standard error of the project's estimated effect, and thus may well lead to an erroneous finding that the effect is statistically significant. For an illustrative example of this mistake, see Institute of Education Sciences, *Identifying and Implementing Educational Interventions Supported By Rigorous Evidence: A User-Friendly Guide*, op. cit., no. 8, page 8.

[12] For a useful discussion of subgroup analyses in group-randomized trials, and why some are feasible and some are not, see Howard S. Bloom, op. cit., no. 2, pp. 35-41.

[13] For an example of a study which translated a program's effect on test scores into grade-level equivalents, see Decker, Paul T., Daniel P. Mayer, and Steven Glazerman, *The Effects of Teach for America on Students: Findings from a National Evaluation,* Mathematica Policy Research, June 9, 2004, at http://www.mathematica-mpr.com/publications/pdfs/teach.pdf.

[14] The Consort Statement at http://www.consort-statement.org/ contains an authoritative, standardized format for reporting these study data, so as to give readers a clear picture of the progress of all participants over the course of the study.