Review of Final Year MSP Evaluations, Performance Period 2007

Analytic and Technical Support for Mathematics and Science Partnerships

Contract # ED-04-CO-0015
Task Order # 0012

March 1, 2010

Prepared for
Patricia O'Connell Johnson
Miriam Lund
Jimmy Yun
Michelle Meier

U.S. Department of Education
OESE/Mathematics and Science Partnerships
400 Maryland Ave, SW
Washington, DC 20208

Prepared by
Ellen Bobronnikov
Hilary Rhodes
Cay Bradley

Review of Final Year MSP Evaluations
Performance Period 2007

OVERVIEW

Abt Associates has been providing evaluation and technical assistance to the U.S. Department of Education's Mathematics and Science Partnership (MSP) Program and its projects since 2005. As part of this support, we look across the portfolio of projects funded by the MSP program to draw lessons on best practices. The current activity was a review of the final evaluations conducted on MSP projects. The purpose of this work was to investigate what could be learned about rigorous evaluation of MSP projects through an analysis of impact evaluations that are being conducted in MSP projects in their final year of funding.

This document presents the findings from our review of the evaluations that were completed by MSP projects in their final year. We conducted in-depth analyses to assess the extent to which the project evaluations met specified criteria. In this report, we discuss the rigor of these evaluations, discuss the challenges and successes in meeting specific evaluation criterion, and make recommendations that may help improve future MSP project evaluations.

When reviewing the MSP project evaluations, we focused primarily on the information contained in the final evaluation reports. We supplemented this information with the evaluation data in the annual performance reports (APRs), as well as information provided directly by projects, in an attempt to fill in missing information and to verify consistent reporting of measures.

The review process proceeded in two stages:

1. Defining the set of projects for review, by first identifying the projects that were in their last year of funding and then selecting projects whose evaluations met specific criteria for inclusion; and
2. Assessing and scoring of project evaluations against a rubric to assess data quality and rigor of implementation of the evaluation.

Each of these stages is described below.

IDENTIFYING THE SET OF PROJECT EVALUATIONS

The first step in this review was to identify the projects whose evaluations would be considered in the review. We limited our review limited to the MSP projects that were in their final year in Performance Period 2007 (PP07): of the 575 projects funded in PP07, 183 reported being in their final year. Because the purpose of this review was to learn about the rigorous impact evaluations that projects conducted, we limited our discussion to those projects that used a research design appropriate for testing the impact of an intervention.[1] Thus, we narrowed the set of projects to

[1] For more information on selecting a design that will provide rigorous evidence of effectiveness, see *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide,* US Department of Education Institute for Education Sciences, National Center for

those that reported implementing an evaluation that used an experimental design, also known as a randomized control trial (RCT) (i.e., where teachers, classrooms, or schools are randomly assigned to a treatment or control group), or a quasi-experimental (QED) design (i.e., where teachers, classrooms, or schools are assigned to a treatment or control group by some method other than random assignment). This reduced the set to 63 projects, which became the focus of our initial review.

After examining the details of the evaluation designs for these 63 projects, we further limited the set to the 37 MSP project evaluations that indeed implemented an experimental or quasi-experimental design with a comparison group and provided sufficient data from both groups to review their evaluations. In this step, we excluded some projects because they did not provide sufficient detail about their evaluations[2], and others because their designs did not include an appropriate comparison group. For example, some projects evaluated pre- and post-test scores for only a treatment group, or compared treatment group scores to established benchmarks. The remainder of our discussion focuses on what we learned from reviewing the evaluations of these 37 projects.

Most evaluations of MSP projects included multiple evaluations of various outcomes. In our review, we considered outcomes of teacher content knowledge, teacher practices, and/or student achievement, and evaluated each as an independent evaluation. Across the final set of 37 projects, 64 unique evaluations were identified. The majority of the evaluations looked at student achievement (55 percent), followed by teacher content knowledge (27 percent), and classroom practices (19 percent).

Our assessment of the rigor of these 64 evaluations follows.

ASSESSING MSP EVALUATIONS FOR RIGOR

We reviewed the information available about each of the 64 evaluations to determine the strength of design and implementation. We used the *Criteria for Classifying Designs of MSP Evaluations* (hereafter referred to as the rubric) that was developed by Westat as part of the Data Quality Initiative (DQI) at the Institute for Education Sciences (IES) within the U.S. Department of Education (see Appendix A). This rubric identifies six criteria for assessing whether the MSP evaluations are conducted in a rigorous manner as follows:

- Baseline equivalence of groups;
- Adequate sample size;
- Use of valid and reliable (or sufficiently tested) measurement instruments;
- Use of consistent methods, procedures, and time frames to collect key outcome data from the treatment and comparison groups;

Education Evaluation and Regional Assistance, December 2003. Retrieved October 23, 2009, from http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf.

[2] Projects that were missing individual data elements were contacted for additional information, but projects that were not able to provide data for the comparison group, or that provided insufficient information to determine the overall design, could not be included in our review.

- Sufficient response and retention rates; and
- Reports of relevant statistics and their statistical significance.

To pass the rubric, evaluations must satisfy the requirements of each criterion. Of the 64 evaluations, five evaluations from four projects successfully met all of the rubric's criteria. One project successfully implemented a rigorous experimental design and three projects successfully implemented evaluations with rigorous quasi-experimental designs.

Since the rubric was developed and approved the rubric *after* the PP07 projects had already designed their evaluations, it is not surprising that a large number of evaluations failed to meet the rubric's criteria. Of greater value than the passing rate, however, are the insights that come from the identification of issues that frequently prevented projects from passing. These insights have led to our recommendations that can help future projects avoid and/or overcome these hurdles and increase the rigor of their evaluation designs so that they produce credible evidence of the MSPs' impact

In the review that follows, we discuss the MSP evaluations' performance on each of the rubric's six criteria. For each, we present information on: (1) how the criterion is defined; (2) the justification for its inclusion; (3) the requirements for passing; (4) results on passing rates among the set of evaluations reviewed; (5) common issues found; and (6) recommendations for meeting the criterion in future evaluations.

CRITERION 1: BASELINE EQUIVALENCE

Description: No significant pre-intervention differences exist between treatment and comparison group participants on variables related to key outcomes, or groups have similar background characteristics.

Justification: Findings from quasi-experimental studies in which baseline equivalence of groups has been demonstrated (or difference has been controlled for in the analysis) are considered more rigorous, as at least some possible alternative explanations for the differences between groups are addressed.

Screening requirements: Evaluations pass the baseline equivalence criterion when their evaluation design meets at least one of the following three conditions:

> 1.1 – Uses an experimental design (i.e., random assignment) that should yield probabilistically equivalent groups and therefore is not required to demonstrate baseline equivalence.
>
> 1.2 – Uses a quasi-experimental design and test for and finds no statistically significant pre-intervention differences between groups on variables related to key outcomes.
>
> 1.3 – Uses a quasi-experimental design and controls for baseline differences in the analysis.

Results: Overall, 25 of the 64 evaluations (39 percent) passed the baseline equivalence criterion. Among the evaluations that passed, 1 evaluation used an experimental design and thus was not required to demonstrate baseline equivalence, and 7 evaluations (28 percent) used statistical testing (e.g., t-tests or chi-square) to demonstrate that there were no significant differences between the groups at the project's start. See Exhibit 1.

Exhibit 1
Percent of Evaluations that Pass Baseline Equivalence Criterion, by Condition

Conditions for Passing	Number (Percent) of Passing Evaluations (N=25)
1.1 Experimental study not required to demonstrate equivalence	1 (4%)
1.2 Quasi-experimental study demonstrating no statistically significant pre-intervention differences between groups on variables related to key outcomes	7 (28)
1.3 Quasi-experimental study addressing baseline differences in analysis (e.g., ANCOVA, inclusion of pretest as covariate, gain score analysis)	17 (68)

Sources: Final evaluation reports, annual performance reports, and related documents

The remaining 7 evaluations (68 percent) that passed the equivalence criterion adjusted for differences at baseline in their analysis. Projects used the following methods to adjust for pretest differences: analyses of variance (ANOVA), analysis of co-variance (ANCOVA) with time and/or pretest results as covariates, and gain score analysis, where the treatment effect is estimated by comparing the mean gain in the treatment group with the mean gain in the comparison group.

Common issues: 39 of the 64 evaluations (61 percent) of the evaluations did not pass the baseline equivalence criterion. The two most common reasons evaluations did not pass are: 1) baseline characteristics are reported without reporting a statistical test for differences, and 2) information critical for complete assessment of baseline equivalence, including sample size and standard deviations, is missing.

Recommendations:

1. Report key characteristics that are associated with outcomes for each group, such as pretest scores and teaching experience. Always include sample sizes when reporting statistics.
2. Test for group mean differences on key characteristics with the appropriate statistical test (e.g. chi-square for dichotomous characteristics, t-test for continuous characteristics). Report the test statistics, such as t-statistic and p-values.

CRITERION 2: SAMPLE SIZE

Description: Sample size is adequate based on a power analysis or on meeting predetermined thresholds for the number of students, teachers, or schools needed to have adequate power.

Justification: Sufficient sample size is needed to build confidence in the results. When calculating adequate sample sizes, the standard practice is to use a significance level of .05 and power (i.e., the probability of detecting an actual difference if it exists) of .80 to estimate an appropriate sample size.

Screening requirements: An evaluation passes if we could confirm that the evaluation's sample size for the evaluation was adequate, that is, when there was sufficient sample size at the level of assignment or analysis.

Results: Just over half of the 64 evaluations (33 evaluations, 52 percent) had adequate sample sizes to detect differences in the outcomes measured. While no evaluation reported using a power analysis, the passing evaluations met or exceeded the threshold sample sizes.

Common issues: 31 of the 64 evaluations did not satisfy the sample size criterion. The two most common reasons that evaluations did not pass this criterion were that sample sizes were not reported or that there was inconsistent reporting of sample sizes within or across project APRs and evaluation reports.

Recommendations:

1. Conduct a power analysis at the design stage of an evaluation to ensure that the study will have a large enough sample, and report the calculations of the power analysis.
2. If you do not conduct a power analysis for the project, ensure you have more than the minimum thresholds noted below.
    - *Teacher outcomes*: 12 schools (for school- or district-level interventions) or 60 teachers (for teacher- or classroom-level interventions)
    - *Student outcomes:* 12 schools (for school- or district-level interventions) or 18 teachers (for teacher- or classroom-level interventions) or 130 students (for student-level interventions)
3. Always provide clear reporting of samples sizes for all groups and subgroups.

CRITERION 3: QUALITY OF THE MEASUREMENT INSTRUMENTS

Description: Quality of measures is demonstrated through use of: existing data collection instruments that have already been deemed valid and reliable to measure key outcomes; data collection instruments developed specifically for the study that are sufficiently pretested; or data collection instruments composed of items from a validated and reliable instrument(s).

Justification: Evaluations need to use instruments that accurately capture the intended outcomes for a group similar to the one being included in the study.

Screening requirements: All instruments used to measure outcomes must have face validity, that is, they must appear to measure what they purport to assess. In addition, the instrument used should be deemed valid and reliable.

Results: 52 of the 64 evaluations (81 percent) were measured with an appropriate instrument. Among the 52 evaluations that passed, 42 (81 percent) were measured using an existing instrument in its entirety. See Exhibit 2. Seven evaluations (13 percent) created a new assessment using items from existing instruments that have been validated and deemed reliable; five evaluations (9 percent) used a full scale from an existing instrument, that is, the full subset of items (e.g., all geometry questions from a mathematics test); and two evaluations (four percent) used selected items from existing instruments. Finally, completely new instruments were developed and validated for the remaining three evaluations (6 percent) that passed this criterion.

Exhibit 2
Percent of Evaluations that Pass Quality of Measurement Instrument Criterion, by Instrument Creation Method

Instrument Creation Method	Number (Percent) of Passing Evaluations (N=52)
Used full existing instrument	42 (81%)
Used full scale from existing instrument(s)	5 (9)
Used items selected from existing instrument(s)	2 (4)
Created all items	3 (6)

Sources: Final evaluation reports, annual performance reports, and related documents

Common issues: 12 of the 64 evaluations (19 percent) did not pass this criterion. This was primarily due to projects not reporting the validity or reliability of the instruments they used.

Recommendations:

1. Use instruments that have been shown to have accurate and consistent scores (i.e., have demonstrated reliability and validity). Where possible, use instruments whose scores have demonstrated reliability and validity for a population similar to the population of your study.

2. If you are creating an assessment for the project, assess and report validity and reliability of scores with a population similar to your respondents. For example, if the focus of your project is upper elementary school teachers, you might also have 5th grade teachers in a school not participating in your program complete the assessment.
3. When selecting items from an existing measure:
    a. Describe previous work that demonstrates that the scores are valid and reliable with a population similar to yours;
    b. Provide references to the manual or other studies discussing the validity and reliability of scores; and
    c. Use full subscales rather than choosing items from across subscales where possible.

CRITERION 4: QUALITY OF THE DATA COLLECTION METHODS

Description: The methods, procedures, and time frames used to collect the key outcome data from treatment and comparison groups are the same or similar enough to limit the possibility of observed differences being attributed to another factor.

Justification: Using consistent methods and procedures and collecting data within a similar time frame helps to ensure that observed differences are not attributable to the passage of time or to differences in testing conditions.

Screening requirements: Evaluations pass the data collection methods criterion if evaluators used the same methods, procedures, and time frame to collect data from the treatment and comparison groups. Since most projects did not specify the data collection procedures used for both groups, if there was no reason to believe there were differences, evaluations were given the benefit of the doubt on this criterion.

Results: 57 of the 64 evaluations (89 percent) reviewed passed the data collection methods criterion

Common issues: Seven of the 64 evaluations (11 percent) did not pass the data collection methods criterion. Documents from these projects suggested, or indicated, that data were collected at different times or that data were not collected systematically for the two groups. Most projects provided little information about data collection or only described the process used with the treatment group.

Recommendations:

1. Make every attempt to collect data from both the treatment and comparison groups for every evaluation. If data cannot be collected from all members of both groups for resource reasons, consider randomly selecting a subset of respondents from both the treatment and control group. For example, if the project can support classroom observations of 20 teachers, select 10 from the treatment group and 10 from the comparison group.
2. Describe and document the data collection procedures.

CRITERION 5: DATA REDUCTION RATES

Description: Key outcomes at the posttest are measured for at least 70 percent of the original sample (treatment and comparison groups combined). In addition, where there is differential attrition of more than 15 percentage points between groups, this difference is accounted for in the statistical analysis.

Justification: Significant sample attrition can bias results, since the participants who drop out of the study may differ from those who remain. It is also important to consider the differential attrition between the treatment and control groups, which can create systematic differences between the groups.

Screening requirements: To pass, the evaluation must meet one of the three conditions described below:

> 5.1 – Posttest data for 70 percent of original sample AND less than 15 percent difference in retained sample between treatment and control groups.
>
> 5.2 **–** Sufficient steps have been taken in the statistical analysis to address the difference.
>
> 5.3 **–** There is evidence that attrition is unrelated to the intervention.

When attrition rates were not provided in the evaluation, where we could we calculated attrition rates by subtracting the posttest N from the pretest N and dividing by the pretest N.

Results: 19 of the 64 evaluations (30 percent) passed the data reduction rates criterion. See Exhibit 3. Eighteen of the evaluations that passed (95 percent) did so because they reported posttest data for at least 70 percent of the original sample and had a difference of less than 15 percentage points in the attrition rates for the treatment and control groups. Two evaluations (11 percent) provided evidence that the sample's attrition was not related to the intervention. No evaluations took significant steps to adjust for any difference during the statistical analysis.

Exhibit 3
Percent of Evaluations that Pass Data Reduction Rates Criterion, by Condition

Conditions for Passing	Number (Percent) of Passing Evaluations (N=19)
5.1 Posttest data for 70% of original sample AND less than 15 percent difference in retained sample between treatment and control groups	18 (95%)
5.2 Sufficient steps have been taken in the statistical analysis to address the difference	0 (0)
5.3 There is evidence that attrition is unrelated to the intervention	2 (11)
Passed More Rigorous Criteria	
Reported consistent sample sizes pre and post intervention, or changes to the full sample could be accounted for	3 (16)
Attrition rates of less than or equal to 20 percent	2 (11)
Differential attrition between groups of less than or equal to 10 percent	1 (5)

Sources: Final evaluation reports, annual performance reports, and related documents
Note: Percents may total more than 100 percent because evaluations could meet multiple criteria.

As indicated in Exhibit 3, some evaluations passed and achieved a more rigorous standard for this criterion: three evaluations reported consistent sample sizes pre and post intervention, or changes to the full sample could always be accounted for; two evaluations had attrition rates of less than or equal to 20 percent; and one evaluation had differential attrition between the groups of less than or equal to 10 percentage points.

Common issues: 45 of the 64 evaluations did not pass the data reduction criterion. Most of these evaluations failed because of missing information. Most commonly, evaluations did not report initial sample sizes for both the treatment and comparison groups, so that attrition rates could not be calculated.

Recommendations:

1. Identify the unit of assignment (unit at which groups were created) and unit of analysis (unit at which outcomes are measured and analyzed).
2. Report the number of units of assignment and units of analysis at the beginning and end of the study.
3. If reporting on subgroups, report sample sizes for all subgroups.

CRITERION 6: RELEVANT STATISTICS REPORTED

Description: Final report includes treatment and comparison group posttest means and tests of statistical significance for key outcomes *or* provides sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation/ standard error).

Justification: Reporting relevant statistics provides critical context for interpreting the reported outcomes and indicates where an observed difference is larger than what would likely be created by chance.

Screening requirements: An evaluation passes if either of the following conditions is met:

> 6.1 – Posttest means and test of significance for key outcomes are included in the evaluation.
>
> 6.2 – Evaluation provides sufficient information to calculate statistical significance (e.g., reports of mean, sample size, standard deviations/standard error).

Results: 24 of the 64 evaluations (38 percent) passed the relevant statistics reported criterion. See Exhibit 4. Fifteen evaluations (63 percent) passed because they presented the statistical significance tests, and ten (42 percent) passed because they provided sufficient information to support statistical testing for group differences.

Several passed and achieved a more rigorous standard than the rubric's threshold: three evaluations matched the unit of assignment with the unit of analysis or made the appropriate adjustments; four evaluations adjusted for pretest differences in the analysis; three evaluations included other covariates in the analysis; and three evaluations reported means, standard deviations, and the number of clusters, where appropriate.

Exhibit 4
Percent of Evaluations that Pass Relevant Statistics Criterion, by Condition

Conditions for Passing	Number (Percent) of Passing Evaluations (N=24)
6.1 Posttest means and test of significance for key outcomes are included in the evaluation	15 (63%)
6.2 Evaluations provide sufficient information to calculate statistical significance	10 (42)
Passed More Rigorous Criteria	
Matched the unit of assignment with the unit of analysis or made the appropriate adjustments	3 (13)
Adjusted for pretest differences in the analysis	4 (17)
Included other covariates in the analysis	3 (13)
Reported means, standard deviations, and the number of clusters, where appropriate	3 (13)

Sources: Final evaluation reports, annual performance reports, and related documents
Note: Percents may total more than 100 percent because evaluations could meet multiple criteria.

Common issues: Forty of the 64 evaluations did not report all relevant statistics. As is the case with the earlier criteria, missing information made it difficult to assess evaluations for statistical significance analysis. Common missing information included reports of means, standard deviations or standard errors, and sample size.

Recommendations:

1. For each evaluation, report mean, standard deviation (or error), and sample size. If reporting a regression model or ANOVA analysis, report the model as usual as well as the mean and standard deviation (or error).
2. Report appropriate test for differences between groups (e.g., t-statistic and p-value if continuous outcome).

SUMMARY

Thirty-seven projects examined project outcomes with either an experimental or quasi-experimental design. Across these projects, we evaluated 64 independent evaluations against each of the six rubric criteria. Five individual evaluations across four projects successfully met all of the rubric's criteria.

Exhibit 5 summarizes the percent of evaluations passing each criterion in the rubric. Evaluations were most likely to meet the criterion regarding the quality of data collection methods (57 evaluations, 89 percent) and least likely to pass the data reduction rate criterion (19 evaluations, 30 percent).

One common issue found across all criteria is that projects did not provide sufficient evidence for us to determine whether they should pass. A simple first step in improving the passing rates of project evaluations would be to report the key data points that help describe the quality of an evaluation. This includes reporting both the initial and ending sample sizes and key statistics such as means, standard deviations, and test statistics (e.g., t-statistic and p-value) for both the treatment and comparison groups across key outcomes.

Exhibit 5
Review of Final Year MSP Evaluations, Performance Period 2007

Criterion	Number (Percent) of Passing Evaluations (N=64)	Key Recommendations
1. Baseline Equivalence	25 (39%)	Complete and report pretesting for differences between groups on key outcomes. Provide full information (e.g., sample, size, mean, standard deviation, test results such as t-statistic) about the pretest.
2. Sample Size	33 (52)	Clearly report sample sizes for all groups and subgroups for key outcomes.
3. Quality of Measurement Instruments	52 (81)	Note the validity and reliability of all instruments used. Use full sub-scales when taking items from existing instruments, where possible. Test for validity and reliability when creating a new instrument.
4. Quality of the Data Collection Methods	57 (89)	Collect data from the comparison and treatment groups at the same time in a systematic fashion.
5. Data Reduction Rates	19 (30)	Report initial and final sample sizes for all groups and subgroups. Note the number of students in the classrooms and the number of students who transfer in and out over the course of the evaluation.
6. Relevant Statistics Reported	24 (38)	Describe the sample sizes, the means, and the standard deviations as well as the statistical tests, used to analyze results.

Sources: Final evaluation reports, annual performance reports, and related documents

Appendix A: Criteria for Classifying Designs of MSP Evaluations

This appendix includes the *Criteria for Classifying Designs of MSP Evaluations* used to determine the number of projects that successfully conducted rigorous evaluations. The criteria were developed by Westat as part of the Data Quality Initiative (DQI) through the Institute for Education Sciences (IES) at the U.S. Department of Education.

Criteria for Classifying Designs of MSP Evaluations

- ❒ **Experimental study—**the study measures the intervention's effect by randomly assigning individuals (or other units, such as classrooms or schools) to a group that participated in the intervention, or to a control group that did not; and then compares post-intervention outcomes for the two groups.

- ❒ **Quasi-experimental study—**the study measures the intervention's effect by comparing post-intervention outcomes for treatment participants with outcomes for a comparison group (that was not exposed to the intervention), chosen through methods other than random assignment. For example:

  - *Comparison-group study with equating*—a study in which statistical controls and/or matching techniques are used to make the treatment and comparison groups similar in their preintervention characteristics.

  - *Regression-discontinuity study*—a study in which individuals (or other units, such as classrooms or schools) are assigned to treatment or comparison groups on the basis of a "cutoff" score on a preintervention nondichotomous measure.

- ❒ **Other**

  - The study uses a design other than a randomized controlled trial, comparison-group study with equating, or regression-discontinuity study, including *pre-post* studies, which measure the intervention's effect based on the pretest to posttest differences of a single group, and comparison-group studies without equating, or nonexperimental studies that compare outcomes of groups that vary with respect to implementation fidelity or program dosage.

Criteria for Assessing Whether Designs Were Conducted Successfully and Yielded Scientifically Valid Results

A. Sample size[3]

- ❒ **Met the criterion**—sample size was adequate (i.e., based on power analysis with recommended significance level=0.05, power=0.8, and a minimum detectable effect informed by the literature or otherwise justified).

- ❒ **Did not meet the criterion** —the sample size was too small.

- ❒ **Did not address the criterion.**

B. Quality of the Measurement Instruments

- ❒ **Met the criterion**—the study used existing data collection instruments that had already been deemed valid and reliable to measure key outcomes; or data collection instruments developed specifically for the study were sufficiently pre-tested with subjects who were comparable to the study sample.

- ❒ **Did not meet the criterion** —the key data collection instruments used in the evaluation lacked evidence of validity and reliability.

- ❒ **Did not address the criterion.**

C. Quality of the Data Collection Methods

- ❒ **Met the criterion**—the methods, procedures, and timeframes used to collect the key outcome data from treatment and control groups were the same.

- ❒ **Did not meet the criterion**—instruments/assessments were administered differently in manner and/or at different times to treatment and control group participants.

[3] Experimental designs were not required to meet this criterion.

D. Data Reduction Rates (i.e., Attrition Rates, Response Rates)

- ❒ **Met the criterion**—(1) the study measured the key outcome variable(s) in the post-tests for at least 70% of the original study sample (treatment and control groups combined) or there is evidence that the high rates of data reduction were unrelated to the intervention, ***and*** (2) the proportion of the original study sample that was retained in follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was similar for both the treatment and control groups (i.e., less or equal to a 15-percent difference), or the proportion of the original study sample that was retained in the follow-up data collection was different for the treatment and control groups, but sufficient steps were taken to address this differential attrition in the statistical analysis.

- ❒ **Did not meet the criterion**—(1) the study failed to measure the key outcome variable(s) in the post-tests for 30% or more of the original study sample (treatment and control groups combined), and there is no evidence that the high rates of data reduction were unrelated to the intervention; ***or*** (2) the proportion of study participants who participated in follow-up data collection activities (e.g., post-intervention surveys) and/or for whom post-intervention data were provided (e.g., test scores) was significantly different for the treatment and control groups (i.e., more than a 15-percent difference) and sufficient steps to address differential attrition were not taken in the statistical analysis.

- ❒ **Did not address the criterion.**

E. Relevant Statistics Reported

- ❒ **Met the criterion**—the final report includes treatment and control group post-test means, and tests of statistical significance for key outcomes; or provides sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation/standard error).

- ❒ **Did not meet the criterion**—the final report does not include treatment and control group post-test means, and/or tests of statistical significance for key outcomes; or provide sufficient information for calculation of statistical significance (e.g., mean, sample size, standard deviation/standard error).

- ❒ **Did not address the criterion.**